OUTLOOK
Urban health & well-being

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

Beating Ebola

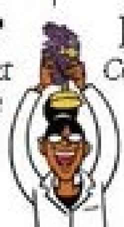
Broad-spectrum antiviral GS-5734 active in monkeys and now progressing through clinical trials **PAGE 381**

NEUROPATHOLOGY

THE SEEDS OF ALZHEIMER'S?

The controversial idea that dementia is transferrable

PAGE 294



REPRODUCIBILITY

INDUSTRIAL REVOLUTION

Commercial partnerships get reliable results

PAGE 299

SYNTHETIC BIOLOGY

ENGINEERING THE FUTURE

Cellular factories for drugs, food and materials

PAGE 401

NATURE.COM/NATURE
17 March 2016 £10
Vol 531, No. 7594

CITY SCIENCE Urban research gets a second chance to impress **p.276**

WORLD VIEW Scientists must share blame for rise of Trump **p.277**

EUREKA! Buoyant beetles drag their feet to fly **p.279**

Africa's elite

A new forum promises to bring deserved prestige to outstanding African researchers, and demonstrates the continent's untapped potential.

How can architects and town planners help clinicians to tackle tuberculosis? What is space-time? These are among the questions being explored by African scientists who last week joined together to open the world's first truly pan-African scientific gathering.

The Next Einstein Forum (NEF), held in Dakar, Senegal, deserves to become a regular feature of the global science landscape. Its purpose: to publicly celebrate and support some of the most outstanding young researchers active in, or closely tied to, the continent.

There were 15 NEF Fellows in Dakar last week — from Egypt, Morocco, South Africa, Senegal, Ethiopia and places in between (see http://nef.org/nef-fellows). All have proved themselves as scientists, and all have deep connections with Africa. Most work on the continent; others are based in prestigious institutions in the United States and in Europe, forming part of the sort of diaspora on which other regions — especially China — have built a thriving research ecosystem.

The launch included a scintillating show by African singers and dancers, and political messages of science-centred ambition by President Macky Sall of Senegal and Rwandan President Paul Kagame. But from the outset, as seasoned conference attendees agreed, it was the young scientists who were the freshest and most compelling feature of the event.

They presented work on a three-wheeled tractor that can negotiate the muddy tracks that make up many of Africa's rural roads, and can thresh maize (corn) and pump water for irrigation; and a theoretical model for the dark energy that drives the Universe's accelerating expansion. They are researching new preventive treatments for cardiovascular diseases that are highly prevalent in black populations, and the fundamentals of semantic data-analysis. They are working towards databases that use artificial intelligence to generate their own hypotheses. And more.

The meeting was not just a showcase of what is excellent. The all-too-familiar challenges faced by Africa's researchers were thoroughly rehearsed, both by the Fellows and by African, European and US policy grandees who sat alongside them on panels.

Neighbouring countries face drastically different challenges and have wildly unequal resources. There is inadequate public and research infrastructure and a lack of intellectual-property protection. A culture that holds that 'science and maths are only for the best' too often hampers teachers who wish to encourage science, and widespread assumptions that these subjects are only for boys dissuade girls from pursuing science, and present discriminatory obstacles to female researchers.

Despite the challenges, it was striking how the meeting was an expression of determination to find the gems among Africa's researchers and support them. Particularly welcome was the sense that fundamental research is as essential as work that has direct societal applications.

A capacity for fundamental science and mathematics is essential if ideas and techniques developed elsewhere are to be adapted and absorbed in African contexts. As Senegal's research minister rightly emphasized, World Health Organization protocols for hepatitis B vaccination in Africa, originally derived in Asian settings, were changed because of feedback from African biologists. And innovations in handling partial differential equations need to be applied not only to fundamental physics, but also to water management.

How to build on what was undoubtedly a successful exhibition of talent? The prime backers of the meeting are the African Institute for Mathematical Sciences — now spreading its wings in several countries from its roots in South Africa (see go.nature.com/9putdt) — and the Robert Bosch Foundation, based in Stuttgart, Germany. Both deserve credit, and would do well to ensure that the Fellows grow in number and that alumni keep coming to the forum as active members of a quasi-family. If that happens, the meeting could develop into a prestige event for those inside and outside Africa who want to understand and support the best of indigenous African research. Credit to Rwanda for hosting the next NEF in 2018.

"The meeting was an expression of determination to find the gems among Africa's researchers."

There is also an opportunity to make the most of the growing number of schemes that support younger scientists in Africa, alongside the NEF — the National Young Academies and Global Young Academy, the Africa Science Leadership Programme at the University of Pretoria (see go.nature.com/fkaq9f) and the DELTAS Africa programme of the Wellcome Trust (see go.nature.com/b23xux).

The drive from the congress centre to the participants' hotels highlighted the disparities between the rich worlds of many attendees' home countries and the streets of Dakar. Poverty and inequality can be reduced only step by step; the step represented by this forum was significant. It showed what powerful commitment there is to be tapped in this emerging generation of young African scientists. The venture and the researchers it represents deserve strong support. ■

Metropolis now

Growing urbanization is heralding a new era of science in the city.

Late last year, Chinese officials reactivated a machine of the state that had lain idle for almost four decades. The government reconvened its Central Urban Work Conference and gave it a crucial task — to report back on how to revamp and revitalize the nation's growing, and choking, cities.

When the expert group issued its recommendations last month, it backtracked on many of the country's previous urban policies that had prized growth above all else. The new plan promises denser streets, to

break the damaging reign of the car, mixed-use neighbourhoods with greater diversity, and more investment in public transport. China has long been said to have 'a thousand cities with the same face'. Now it is trying to put a smile on them.

Amid the scientific and social priorities for the coming years, the study and design of cities must be right at the top. Humans are now an urban species and have been since city dwellers started to outnumber rural folk for the first time almost a decade ago. The trend is set to continue, and the United Nations has estimated that 70% of the global population will live in an urban environment by 2050.

What kind of environment will that be? The signs so far are not good. Too many people, especially those in rapidly developing (and urbanizing) countries experience city life much as Charles Dickens put it in his 1850s work *Little Dorrit*: "Miles of close wells and pits of houses, where the inhabitants gasped for air, stretched far away towards every point of the compass. Through the heart of the town a deadly sewer ebbed and flowed, in the place of a fine fresh river."

Scientists are responding. Universities in many of the world's cities — London, New York, Boston, Madrid, Glasgow, Zurich and Singapore among them — are leading a new wave of evidence-driven, data-rich research that aims to understand what makes cities tick, and to keep them running smoothly. Some of these issues, from how best city dwellers should move around, to how to protect their water and them from it, are discussed in a collection of articles this week in a *Nature* Outlook supplement on urban health and well-being (turn to page 306).

Science and technology has a chequered history in the city. On the plus side, great urban visionaries of the past — such as the British town planner Patrick Geddes — trained as scientists, and were able to bring the ideas of ecology and the natural environment to their social tasks. In a less enlightened contribution, some of the haste to design cities around the automobile was justified by claims to rational science. (Even today, many cities in the developing world spend 70% of their transport budgets on serving the car, even though 70% of trips are made by foot or public transport.)

By the 1960s, cities were almost a frontier too far even for science. Asked by then-US President Lyndon Johnson to solve the social problems rooted in US urban areas, a specially convened group of scientists in Woods Hole in Massachusetts responded that "creating a safe, happy city is a greater challenge than a trip to the moon". Residents of the Bronx in New York City would have agreed: a botched attempt to model demand for fire services in the 1970s contributed to a series of ill-judged fire-station closures and an outbreak of (oddly, not predicted) fires.

"Urban science has some way to go to restore its reputation."

Urban science has some way to go to restore its reputation, but the era of big data offers an opportunity, and a new way of thinking. Even as civic leaders crow about the unique merits of their towns, research on cities is trying to dismantle them. Rather than looking at what makes cities different — and then planning accordingly — modern urban science seeks what they have in common.

The new models of urban life start from the ground up and track, for example, people's journeys and the reasons for them, rather than the flow of traffic through a specific, frequently gridlocked roundabout or intersection. They try to make city science quantifiable, testable and reproducible. It's a big ask — but that trip to the Moon was achieved. And as the same Woods Hole scientists also told Johnson, the problems of the city "nevertheless, can be attacked in the same logical way we have gone about exploring the universe". ■

Practical DNA

The promise of DNA origami shows signs of coming to fruition a decade after its debut.

Science seeks to understand the mechanisms of nature, to develop tools of investigation and to make useful and sometimes revolutionary things with which to build our future. And every now and again, a piece of science comes along that seems like a work of art.

All of this was exemplified by a research paper published in *Nature* ten years ago that, literally, produced smiles (see *Nature* **440,** 297–302; 2006). Using an astoundingly simple and general method to assemble strands of DNA into arbitrary shapes, the research generated 'smileys' that graced the cover of *Nature* and announced the arrival of DNA origami to the world.

The robustness of this method changed the game for DNA nanotechnology, which has since developed at an astonishing pace. It is a beautiful demonstration of how science can progress.

The concept behind DNA origami was laid down in the early 1980s by crystallographer Nadrian Seeman, who realized that the ability of DNA molecules to carry and transfer information according to strict base-pairing rules could be used to rationally assemble structures with precisely controlled nanoscale features.

This unprecedented level of programmability makes DNA a unique building material. Nanodesigners have embraced the biomolecule to fabricate intricate tiled patterns, boxes with lids that can be opened and arrays of precisely located binding elements that can incorporate proteins, dyes and other functional materials into regular lattices.

Pivotal to the success of DNA as a nanoscale building material have been automated methods to synthesize short DNA molecules of any sequence. A detailed understanding of how base-pairing translates into the formation of DNA double helices has also been crucial. Such helices control the shapes into which DNA molecules with given sequences will fold.

DNA origami provides the missing ingredient: a versatile yet straightforward assembly method. Computer-aided design programs determine how DNA scaffolds can be folded to realize desired structures, as well as which short DNA strands, or staples, are needed to hold the structures in shape.

Individual structures can also be assembled into more complex patterns, and sites that bind to functional materials can be introduced at any position.

The many eye-catching structures that have been built have pleased those of us with an appreciation of beauty. But even the most creative science will ultimately face the question: what is the point?

DNA nanotechnology has long searched for relevance. It is unrivalled in its ability to build complex structures with near-atomic precision, but the results tend to be labile, soft and so small that it is a challenge to put them to practical use.

Yet applications that address basic problems in science have emerged. DNA structures can serve as tools for determining the structures of proteins or as templates for assembling electronic components and basic devices. Responsive DNA structures can target diseased cells, and artificial membrane channels formed from DNA can act as single-molecule sensors.

Real-world applications might become feasible through recent developments — for example, improvements to the folding process that reduce assembly time and boost yield. Initial steps have also been taken to efficiently pair DNA nanostructures with technologically relevant substrates.

Many challenges remain, and DNA nanotechnology is far from maturity. But a growing number of scientists are entering the field to make more than just art. Watch this space. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

WORLD VIEW *A personal take on events*

CHRIS CLARK/AP PHOTO

The elephant in the room we can't ignore

If Donald Trump were to trigger a crisis in Western democracy, scientists would need to look at their part in its downfall, says **Colin Macilwain**.

The annual meeting of the American Association for the Advancement of Science (AAAS) in Washington DC last month was one of the best I've witnessed in more than 20 years of regular attendance. The policy sessions were packed and genuinely stimulating. I met tons of smart, influential people I hadn't seen for ages, and we all enjoyed a good chinwag about how better to engage with the public — the meeting's theme for 2016.

The only trouble was what was going on outside the hotel — in the United States and the world at large.

In fact, the AAAS meeting took place in a sort of semi-conscious never-never land. The science-policy crowd talked a great game even as the pillars of the republic crashed noisily down around their heads.

Supporters or representatives of Donald Trump, the likely Republican nominee for this November's US presidential election, his extremely conservative rival Ted Cruz, or even Bernie Sanders, the Democrat insurgent, were simply not involved in these discussions. They never are. Senior scientists are instead inextricably linked to the centrist, free-market political establishment that has tended to rule, but which is now falling dangerously from public favour.

It is not just in the United States that this consensus — and perhaps democracy itself — is in danger. Poland has just elected a reactionary government that is clamping down on press freedom; France is toying with electing far-right politician Marine Le Pen to the presidency; and the rest of the world's elected leaders are each threatened, to a greater or lesser extent, by economic and migration crises. Populist nationalism is on the march again — exemplified by the rise of Trump, whose mode of operation does not countenance the opinions, advice or goodwill of anybody else.

POPULIST NATIONALISM IS ON THE MARCH AGAIN — EXEMPLIFIED BY THE RISE OF TRUMP.

Not for nothing are people being urged to read *Rubicon: The Triumph and Tragedy of the Roman Republic*, Tom Holland's summary of how Rome fell. The establishment — with which science has habitually enjoyed a genial, if subservient relationship — is on the rocks.

Many laboratory researchers perceive this, I fear, to be someone else's problem. But it isn't. If the West is really in its decline-and-fall stage, its Caligula stage, its Donald Trump stage, then this isn't just an issue for political and financial elites. It's also a problem for the 'experts' who crawl around after these elites, massaging their egos and defending their interests.

The playwright Bertolt Brecht had a good line on expertise. In his plays, doctors, lawyers and other 'experts' are generally portrayed in threes. They squabble haplessly among themselves, each manoeuvring into the position that most elevates themselves in the eyes of their aristocratic paymaster.

➲ **NATURE.COM**
Discuss this article online at:
go.nature.com/5xjnt5

And that, sadly, is the role to which senior scientific leaders have sometimes reduced themselves. In the main, they have been happy to accept the autocracy of politics and finance, even, like the president of the European Research Council, hanging around at the annual meeting of business leaders at Davos in Switzerland, hoping to pick up crumbs from the rich man's table.

The problem extends down into the community itself. We like to talk about 'engaging the public', but many scientists really just want to talk at them. And too many ordinary scientists hold politicians in utter intellectual contempt — even though it is the scientists who have chosen a career that allows them to pursue relatively simple problems (such as building a machine to detect gravitational waves) rather than genuinely difficult ones (such as running a social-care programme in a small town).

And those senior scientists who do engage with the government or public — as scientific advisers, for example — often take up highly political positions without acknowledging that they are doing so. For example, they support free-trade agreements that cede the right of democratic governments to control things such as cigarette advertising or pesticide use without hard, scientific evidence. This is a political position that is pursued with great dedication by global corporations — and that is haplessly bought into by many scientists without a thought for its consequences.

I admit that it is difficult to bring more subtle and varied political approaches to the table. Groups of researchers that have tried to do so — such as the Federation of American Scientists and Union of Concerned Scientists — have struggled to gain traction. Still, there is a fresher, grassroots movement, exemplified by local 'sceptics' groups, through which younger scientists are trying to make their work relate to society's wider concerns.

But at the top, there is paralysis: leading scientific organizations do little except chase money and reinforce the ruling nexus of politics and finance — even since the financial crisis of 2008, which discredited the free-market philosophy that underpins that nexus. I argued years ago (see *Nature* **479,** 447; 2011) that scientific leaders had failed to respond in any meaningful way to that collapse, and I'm still waiting.

The political structure of the West is in deep trouble, and should it fall apart, there will be plenty of blame to go around. Most will go to political and financial elites, or to rowdy mobs. But some will belong to people in the middle who have taken public funds, defended elites and then stood back and watched as democracy got ridden over a cliff. ■

Colin Macilwain *writes on science and policy from Edinburgh, UK.*
e-mail: cfmworldview@googlemail.com

RESEARCH HIGHLIGHTS

CARDIOVASCULAR BIOLOGY

Gut microbes raise heart-attack risk

Gut microbes produce a chemical that enhances clotting in the arteries, increasing the risk of heart attack and stroke.

Stanley Hazen of the Cleveland Clinic in Ohio and his colleagues treated human platelets, which form blood clots, with a compound called TMAO. This is made in the body from a waste product of gut microbes, and has been linked to heart disease. The team found that TMAO made the platelets form artery-blocking clots faster. The researchers increased blood TMAO levels in mice by feeding them a diet that was rich in choline, a TMAO precursor, and found that the animals formed clots faster than did those with lower TMAO levels.

This effect was not seen in animals that lacked gut microbes or that were treated with antibiotics. When intestinal microbes from mice that produced high levels of TMAO were transplanted into mice with no gut microbes, the recipients' clotting risk increased. The results reveal a link between diet, gut microbes and heart-disease risk, the authors say.

Cell **http://doi.org/bdb2 (2016)**

TAKAO ONOZATO/AFLO/GETTY

ANIMAL BEHAVIOUR

Order of notes is key in bird calls

A bird species derives different meanings from varying combinations of notes, just as humans understand complex meanings from words combined in different ways.

Toshitaka Suzuki of the Graduate University for Advanced Studies in Hayama, Japan, and his colleagues played recordings of four notes — A, B, C and D — in different orders for the Japanese great tit (*Parus minor*; pictured), which normally uses more than ten different notes in its calls. Playing ABC prompted the birds to scan horizontally for predators. On hearing a repeated D note, the birds approached the source of the calls. ABC–D calls elicited both behaviours, but playing D–ABC invoked little or no response.

The authors suggest that the order of the notes determines meaning, and say that this is the first experimental evidence for 'compositional syntax' in a wild animal.

Nature Commun. 7, **10986 (2016)**

CHEMICAL ENGINEERING

Waste gas makes liquid fuel

Waste gases containing carbon dioxide can be converted into diesel, thanks to a bacterium and an engineered yeast.

Gregory Stephanopoulos and his colleagues at the Massachusetts Institute of Technology in Cambridge developed a two-stage process that uses bioreactors to create liquid fuel out of gas mixtures containing CO_2. The first stage involves the bacterium *Moorella thermoacetica*, which converts mixtures of CO_2 and other gases such as carbon monoxide or hydrogen into acetic acid. An engineered yeast, *Yarrowia lipolytica*, then transforms the acetic acid into an oil that can be turned into diesel using existing industrial processes.

This method, with further enhancements to boost efficiency, could be used to produce fuel from the waste gases that are generated by industrial sites such as steel mills and coal-fired power plants, the authors say.

Proc. Natl Acad. Sci. USA **http://doi.org/bdb5 (2016)**

CANCER

Gene blocks anti-tumour response

A common cancer gene works in part by helping tumours to evade immune cells.

Dean Felsher of Stanford University in California and his colleagues studied the effects of *MYC* — a gene that is often overexpressed in cancer — in a mouse model of a type of leukaemia. They found that higher *MYC* expression levels increased the production of two proteins, PD-L1 and CD47, that help cancer cells to hide from the immune system. When *MYC* was inactivated, CD47 and PD-L1 levels dropped and tumour size decreased. Tumour data from humans showed a strong link between levels of *MYC* expression and levels of these immune-evasion signals.

People with cancers that overexpress *MYC* could benefit from treatments that

boost the immune attack against tumours, the authors suggest.
Science **http://doi.org/bc7p (2016)**

NEUROSCIENCE

Altered sensations in anxiety

Anxiety disorders could involve not only cognitive, but also sensory changes in the brain.

Recent studies have suggested that people with anxiety, after learning a negative stimulus, respond negatively to similar but neutral stimuli more often than healthy people. Rony Paz at the Weizmann Institute of Science in Rehovot, Israel, and his colleagues found that individuals with anxiety disorders also perceive these stimuli less precisely than healthy people do. After learning to associate a tone with either monetary gain or loss, participants were asked to decide whether a series of other sounds were a match to the previous ones or were new. People with anxiety disorders mistook a wider range of frequencies for the tones they had learned, compared with healthy people. Learned tones and neighbouring sounds triggered brain activity that showed greater similarity in people with the disorders than in healthy people. This effect was in the brain's auditory cortex and in the amygdala, which processes fear.

The findings suggest that people with anxiety have altered perception of certain stimuli, the authors say.
Curr. Biol. **http://doi.org/bc3z (2016)**

MATTIAS LANAS

MICROBIOLOGY

Bacterial toxins invite infections

Certain pneumonia-causing bacteria produce compounds that help other pathogenic bacteria to spread through the lungs.

Bret Sellman at MedImmune, a biotechnology firm in Gaithersburg, Maryland, and his colleagues infected mice with a variety of bacterial species, either individually or in combination with *Staphylococcus aureus*, which can cause respiratory and other infections. Mice that were co-infected with *S. aureus* had higher levels of both microbe species in their lungs, and were more likely to die than animals infected with a single pathogen. The team found that a protein produced by *S. aureus*, called α-toxin, aids the growth of several bacterial species by impairing immune-cell function. Early treatment with an antibody against α-toxin helped to eliminate *S. aureus* and prevented other pathogens from multiplying.

The authors suggest that antibody-based treatments targeting a single bacterial species could help some people who are infected with multiple pathogens.
Sci. Transl. Med. 8, **329ra31 (2016)**

BIOMECHANICS

How flying beetles waterski

The waterlily beetle exploits properties of the interface where air and water meet to glide quickly across the surface of ponds.

Manu Prakash at Stanford University in California and his colleagues filmed waterlily beetles (*Galerucella nymphaeae*; **pictured**) at 3,000 frames per second to characterize the mechanics of the insect's unusual mode of flight on the two-dimensional surface. They found that the claws on the beetles' legs remain submerged during flight, anchoring the insects to the water.

Keeping four of their six legs on the water, the insects use the fluid's surface tension to support their body weight and move by flapping their hind wings, cruising along the water's surface at speeds of up to half a metre per second. Moving any faster would render them airborne, because the ripples they produced on the water would release their anchors.
J. Exp. Biol. 219, **752–766 (2016)**

ESO/F. CHAR

ASTRONOMY

Milky Way's bulging waistline

The mass of stars in the Milky Way's central bulge (**pictured**) is about 20 billion times the mass of the Sun — a much higher estimate than in most previous studies.

The central bulge protrudes from the Galaxy's main disk like the yolk of a fried egg and hosts a large density of stars. To estimate the mass of those stars, Elena Valenti at the European Southern Observatory (ESO) in Garching, Germany, and her team used a catalogue of stars of a particular type. The data are from ESO's Visible and Infrared Survey Telescope for Astronomy (VISTA) at the Paranal Observatory in Chile. The team also did a deeper study of all the stars in a small region of the bulge, in part using the Hubble Space Telescope.

Valenti says that this is the first study of the bulge based entirely on observation, without the help of theoretical models.
Astron. Astrophys. 587, **L6 (2016)**

LONGEVITY

Genetic switches for long life

Researchers have homed in on the genetic control points that allow nematodes to live longer when they are on a low-calorie diet.

A team led by Jing-Dong Han of the Chinese Academy of Sciences in Shanghai analysed gene-expression changes over time in the nematode *Caenorhabditis elegans*. The worms were subjected either to caloric restriction or to intermittent fasting, both of which extend worm lifespan.

The team identified changes in the expression of various genes at different times, with metabolism genes responding early during the diet, and those linked to cell division and DNA repair changing later on. The researchers then used an algorithm to identify three sets of genes that regulate this altered expression. Changing the activity of the pathways controlled by these genes extended the lifespan of the worms, mimicking the effect of dietary restriction.
Cell Metab. 23, **529–540 (2016)**

SEVEN DAYS *The news in brief*

EVENTS

AlphaGo victorious

The world's leading Go player, South Korea's Lee Sedol, lost his final match in Seoul against Google DeepMind's AlphaGo machine on 15 March. The tightly fought game brought the best-of-five competition to an end with four wins for the computer versus one for the human player. Sedol came back from three consecutive losses to beat the artificial-intelligence system in the fourth match, but ultimately missed out on the US$1-million prize. Go originated more than 2,500 years ago in China and involves placing black and white counters on a board. See page 284 for more.

Brexit warning

Physicist Stephen Hawking is one of more than 150 scientists, mathematicians, economists and engineers at the University of Cambridge, UK, who warn of a disaster for the nation's science if Britain exits the European Union (known as Brexit). A referendum to be held on 23 June will ask whether the country should leave the EU. In a 10 March letter to *The Times*, organized by protein scientist Alan Fersht, the group argues that the free movement of workers between EU countries helps in the recruitment of high-quality researchers to the United Kingdom. The letter's signatories are all fellows of the Royal Society in London.

Zika meeting

With the Zika virus still spreading rapidly across the Americas, the World Health Organization (WHO) in Geneva held an emergency meeting on mosquito control on 14–15 March. The WHO's Vector Control Advisory Group intends to review evidence to support new and innovative techniques for combating the *Aedes aegypti* mosquitoes that transmit Zika virus, along with dengue and Chikungunya viruses. These techniques include deploying mosquitoes that have been made infertile through genetic modification or irradiation.

PRIZES

Fermat proof prize

Andrew Wiles has received the 2016 Abel Prize for mathematics for his solution to Fermat's last theorem, the Norwegian Academy of Science and Letters announced on 15 March. The problem had stumped some of the world's greatest minds for three and a half centuries. Wiles, a number theorist now at the University of Oxford, UK, will receive 6 million kroner (US$700,000) for his 1994 proof showing that there cannot be any positive whole numbers x, y and z such that $x^n + y^n = z^n$, if n is greater than 2. See go.nature.com/yf1nxj for more.

FACILITIES

Infrastructure map

The European Commission has published its latest wish list of the research-infrastructure projects that it considers most deserving of continent-wide support. The European Strategy Forum on Research Infrastructures road map, released on 10 March, details 21 facilities across all scientific areas to help national governments to prioritize how they spend infrastructure money, and

MAGNOLIA PICTURES/COURTESY EVERETT COLLECTION/REX/SHUTTERSTOCK

Famous killer whale nears end of life

Tilikum, a killer whale (*Orcinus orca*) at SeaWorld in Orlando, Florida, has an incurable lung infection, the theme park's veterinary team has announced. In February 2010, Tilikum dragged his trainer Dawn Brancheau into the pool and killed her. The whale was also involved in two deaths in the 1990s, and the story of his life in captivity was told in the controversial 2013 documentary film *Blackfish*. SeaWorld bought Tilikum in 1983; he is thought to be 35 years old. The species' life expectancy in captivity versus that in the wild is still debated by scientists.

NUMBER CRUNCH

10

Consecutive months in which the global monthly temperature record has been broken. February's temperature was 1.35 °C above average for the month. A strong El Niño weather system has contributed to the record-breaking run.

Source: NOAA

to encourage them to share costs and responsibilities. New facilities listed in the 2016 road map include two in environmental sciences and one in health and food sciences, as well as solar and neutrino telescopes and an infrastructure for scientific research into cultural heritage.

POLICY

Call to save bees

The US Government Accountability Office (GAO) says that US regulatory bodies need to do more to protect bee populations. In a report made public on 11 March, the GAO called on the US Department of Agriculture (USDA) to work more closely with other agencies to protect bee health. The report says that although the USDA has upped efforts to monitor honeybee colonies managed by beekeepers, it does not coordinate the monitoring of wild, native bees. The report also recommends that the Environmental Protection Agency identifies the mixtures of pesticides most commonly used by farmers.

PEOPLE

Minister keeps title

WIN MCNAMEE/GETTY

German defence minister Ursula von der Leyen (**pictured**), who was accused in September 2015 of plagiarism in her medical dissertation in obstetrics, will not lose the title of doctor or her job. The senate of Hanover Medical School, which awarded the title in 1990, announced on 9 March that its formal investigation revealed that some passages in von der Leyen's dissertation were copied from original sources. But these were mostly in the introduction, it said, and the main body of research was original and valid. Since 2011, two German federal ministers have lost their titles and government posts to plagiarism charges.

BUSINESS

Gene data shared

Researchers and the public can now access a database of anonymized genetic information from 10,000 people with hereditary breast or ovarian cancer. The database, called AmbryShare, was launched on 8 March by Ambry Genetics, a genetic-testing company in Aliso Viejo, California — making Ambry the first private company to release its customers' information for free. The Broad Institute of MIT and Harvard in Cambridge, Massachusetts, has an open-access database of more than 60,000 genomes collected from the public, but AmbryShare's data currently focus on specific diseases. Ambry hopes to release up to 200,000 aggregated genomes per year from people with various conditions.

India vaccine fight

The medical charity Médecins Sans Frontières (MSF) is challenging pharmaceutical company Pfizer's application for a patent in India on pneumonia vaccine PCV13, marketed as Prevenar 13 in India. MSF says that it wants to allow other manufacturers to make the vaccine, and lower its cost. The 11 March challenge asserts that the method that Pfizer is trying to patent is too obvious to deserve a patent under Indian law. Pfizer is reported as saying that the complexity of the vaccine justifies the price. In partnership with the vaccine alliance GAVI, Pfizer has reduced the price of Prevenar since 2013.

COMING UP

17–18 MARCH
Commercializing 3D printing for biological applications is discussed at the second Tissue Engineering, Biofabrication & 3D-Bioprinting in Life Sciences conference in Boston, Massachusetts.
go.nature.com/rggrat

21–23 MARCH
NASA holds a meeting in Washington DC to develop its technology road maps.
go.nature.com/dhmq2e

21–25 MARCH
The annual Lunar and Planetary Science Conference convenes in The Woodlands, Texas.
go.nature.com/qpnoxd

Mosquito trial

A proposed field trial of genetically modified mosquitoes in the Florida Keys poses no threat to human health or the environment, the US Food and Drug Administration has determined. Members of the public have 30 days to submit comments on the draft assessment, which was released on 11 March. The *Aedes aegypti* mosquitoes developed by Oxitec of Oxford, UK, are engineered to produce short-lived young to temporarily reduce mosquito populations and combat diseases that they carry. The project has received increased attention from the media and politicians amid concerns about the spread of Zika virus.

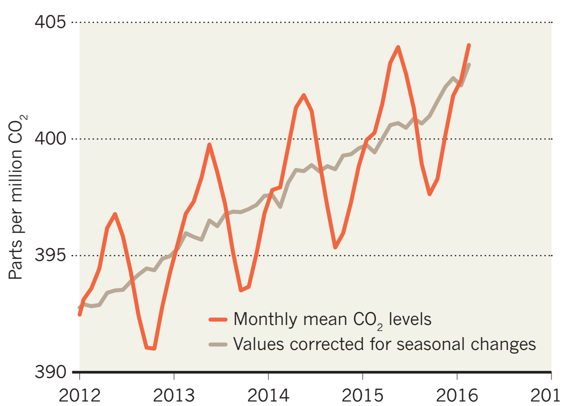
TREND WATCH

SOURCE: NOAA

The level of atmospheric carbon dioxide at the Mauna Loa Observatory in Hawaii rose by 3.05 parts per million (p.p.m.) in 2015 — the largest annual increase since records began 56 years ago, says the US National Oceanic and Atmospheric Administration. After correcting for seasonal swings from plant-growth cycles in the Northern Hemisphere, the average CO_2 concentration in 2015 was 400.83 p.p.m. — a 43% rise compared to the CO_2 level of around 280 p.p.m. that existed during the pre-industrial era.

CARBON DIOXIDE LEVELS BREAK RECORDS

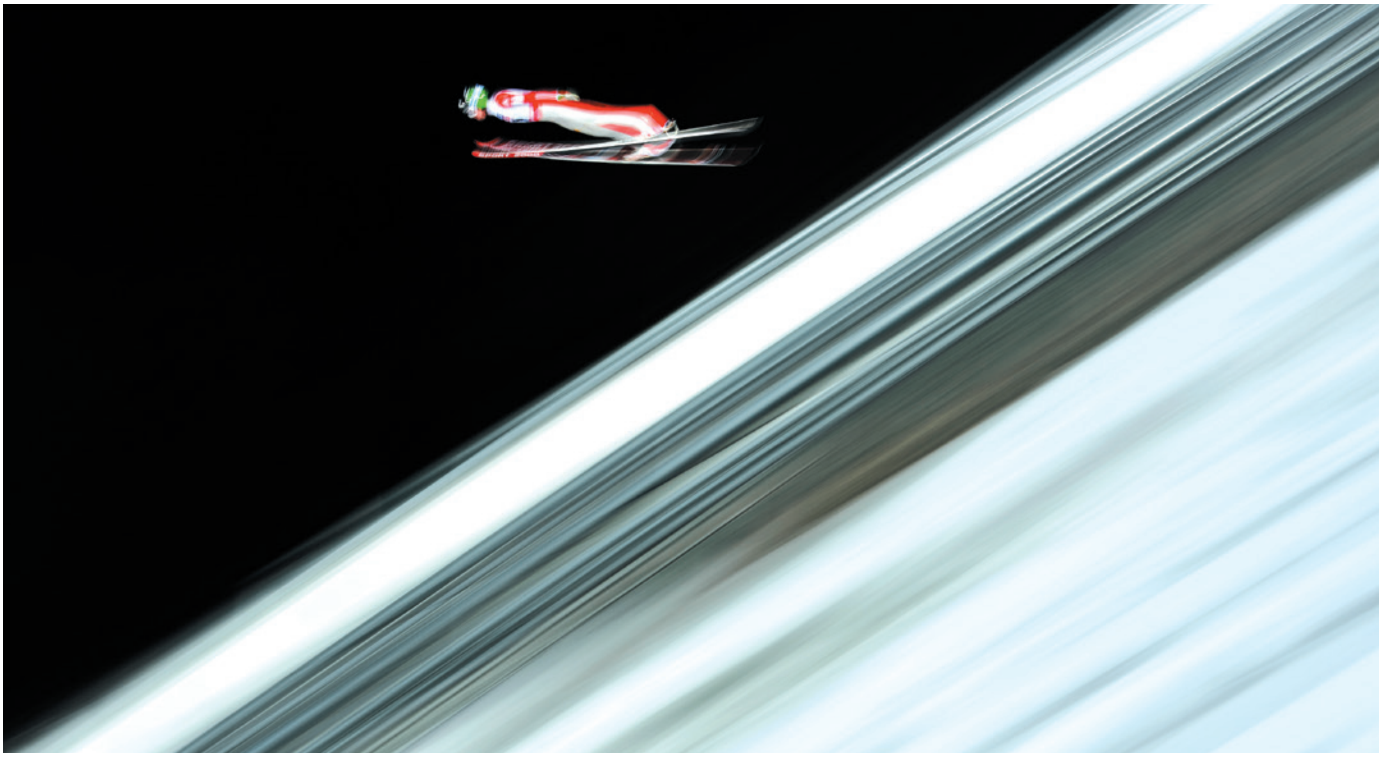
The Mauna Loa Observatory in Hawaii recorded the biggest ever increase in the growth rate of CO_2 concentrations in 2015.



➲ NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

MATTHIAS HANGST/GETTY



Ski jumpers rely on finely tuned skills to glide through the air and land safely.

NEUROSCIENCE

Performance boost paves way for 'brain doping'

Electrical stimulation seems to boost endurance in preliminary studies.

BY SARA REARDON

Elite ski jumpers rely on extreme balance and power to descend the steep slopes that allow them to reach up to 100 kilometres per hour. But the US Ski and Snowboard Association (USSA) is seeking to give its elite athletes an edge by training a different muscle: the mind.

Working with Halo Neuroscience in San Francisco, California, the sports group is testing whether stimulating the brain with electricity can improve the performance of ski jumpers by making it easier for them to hone their skills. Other research suggests that targeted brain stimulation can reduce an athlete's ability to perceive fatigue[1]. Such technologies could aid recovery from injury or let athletes try 'brain doping' to gain a competitive advantage.

Yet many scientists question whether brain stimulation is as effective as its proponents claim, pointing out that studies have looked at only small groups of people. "They're cool findings, but who knows what they mean," says cognitive psychologist Jared Horvath at the University of Melbourne in Australia.

The USSA is working with Halo to judge the efficacy of a device that delivers electricity to the motor cortex, an area of the brain that controls physical skills. The company claims that the stimulation helps the brain to build new connections as it learns a skill. It tested its device in an unpublished study of seven elite Nordic ski jumpers, including Olympic athletes.

Four times per week, for two weeks, the skiers practised jumping onto an unstable platform. Four athletes received transcranial direct-current stimulation (tDCS) as they trained; the other three received a sham procedure. The ▶

stimulation ultimately improved the athletes' jumping force by 70% and their coordination by 80%, compared with the sham group, Halo announced in February.

Troy Taylor, high-performance director for the USSA, is encouraged by the results — but concedes that they are preliminary.

PUSHING THE LIMITS OF ENDURANCE

Another study, presented on 7 March at the Biomedical Basis of Elite Performance meeting in Nottingham, UK, suggests that tDCS may reduce the perception of fatigue. Sports scientist Lex Mauger of the University of Kent in Canterbury, UK, and his colleagues found that stimulating the motor-cortex region that controls leg function allows cyclists to pedal longer without feeling tired.

The researchers stimulated the brains of 12 untrained volunteers before directing the athletes to pedal stationary bicycles until they were exhausted. Every minute, they asked the cyclists to rate their level of effort.

Volunteers who received tDCS were able to pedal two minutes longer, on average, than were those who were given a sham treatment. They also rated themselves as less tired. But there was no difference in heart rate or the lactate level in the muscles between the treatment and control groups. This suggests that changes in brain perception, rather than muscle pain or other body feedback, drove the improved performance.

Alexandre Okano, a biological engineer at Federal University of Rio Grande do Norte in Brazil, found similar increases in cyclists' performance when he stimulated the brain's temporal cortex, which is involved in body awareness and in automatic functions such as breathing[2]. This suggests that the temporal and motor cortices are connected in ways that are not understood, or that tDCS does not target locations in the brain precisely, Okano says.

These results support the notion that the brain manages exertion by collating feedback from the body and then slowing muscles to prevent fatigue, says Dylan Edwards, a neurophysiologist at Burke Medical Research Institute in White Plains, New York[3]. "Even when you think you're exercising as hard as you can, there is always some reserve of ability," he says.

TRICKY TESTS

But Horvath cautions that little is known about the long-term effects of stimulating the brain. And others are sceptical of the technique's potential to increase performance. Vincent Walsh, a neuroscientist at University College London, notes that the methods used in tDCS studies often differ between research groups — and might not always be optimized.

For instance, the fairly intense amount of electricity that Mauger's team used has been shown to sometimes have complex and unintended effects on the brain's activity[4].

Replicating such experiments is difficult because of variations in how people respond to brain stimulation. Some people do not respond at all; others might respond only when stimulated in a certain way. And even an individual's response can differ from day to day. Edwards says that it is important to map these differences if tDCS is to be used therapeutically or for other purposes. "We're moving toward customized prescription of brain stimulation," he says.

Nonetheless, the use of tDCS in sports is only likely to increase. Stimulating the motor cortex, for instance, seems to increase dexterity, so videogamers have been quick to take up the technique. And it is increasingly easy to acquire stimulation devices; Halo has begun to market its equipment for the express purpose of increasing athletic performance.

Taylor compares the use of brain stimulation by athletes to eating carbohydrates ahead of an athletic event, in the hopes of boosting endurance. "It piggybacks on the ability to learn," he says. "It's not introducing something artificial into the body."

But Edwards worries that the availability of tDCS devices will tempt athletes to try "brain doping", in part because there is no way to detect its use. "If this is real," he says, "then absolutely the Olympics should be concerned about it." ■

1. Cogiamanian, F. *et al. Eur. J. Neurosci.* **26,** 242–249 (2007).
2. Okano, A. H. *et al. Br. J. Sports Med.* **49,** 1213–1218 (2015).
3. Noakes, T. D. *Sports Med.* **37,** 374–377 (2007).
4. Batsikadze, G., Moliadze, V., Paulus, W., Kuo, M.-F. & Nitsche, M. A. *J. Physiol.* **591,** 1987–2000 (2013).

ARTIFICIAL INTELLIGENCE

What Google's winning Go algorithm will do next

AlphaGo's techniques could have broad uses, but moving beyond games is a challenge.

BY ELIZABETH GIBNEY

Following the defeat of one of its finest human players, the ancient game of Go has joined the growing list of tasks at which computers perform better than humans. In a 6-day tournament in Seoul, watched by a reported 100 million people around the world, the computer algorithm AlphaGo, created by the Google-owned company DeepMind, beat Go professional Lee Sedol by 4 games to 1. The complexity and intuitive nature of the ancient board game had established Go as one the greatest challenges in artificial intelligence (AI). Now the big question is what the DeepMind team will turn to next.

AlphaGo's general-purpose approach — which was mainly learned, with a few elements crafted specifically for the game — could be applied to problems that involve pattern recognition, decision-making and planning. But the approach is also limited. "It's really impressive, but at the same time, there are still a lot of challenges," says Yoshua Bengio, a computer scientist at the University of Montreal in Canada.

Lee, who had predicted that he would win the Google tournament in a landslide, was shocked by his loss. In October, AlphaGo beat European champion Fan Hui. But the version of the program that won in Seoul is significantly stronger, says Jonathan Schaeffer, a computer scientist at the University of Alberta in Edmonton, Canada, whose Chinook software mastered draughts in 2007: "I expected them to use more computational resources and do a lot more learning, but I still didn't expect to see this amazing level of performance."

The improvement was largely down to the fact that the more AlphaGo plays, the better it gets, says Miles Brundage, a social scientist at Arizona State University in Tempe, who studies trends in AI. AlphaGo uses a brain-inspired architecture known as a neural network, in which connections between layers of simulated neurons strengthen on the basis of experience. It learned by first studying 30 million Go positions from human games and then improving by playing itself over and over again, a technique known as reinforcement learning. Then, DeepMind combined AlphaGo's ability

KIM HONG-JI/TPX/REUTERS

Professional Go player Lee Sedol (centre) after his 4–1 defeat by Google's AlphaGo algorithm.

to recognize successful board configurations with a 'look-ahead search', in which it explores the consequences of playing promising moves and uses that to decide which one to pick.

Next, DeepMind could tackle more games. Most board games, in which players tend to have access to all information about play, are now solved. But machines still cannot beat humans at multiplayer poker, say, in which each player sees only their own cards. The DeepMind team has expressed an interest in tackling Starcraft, a science-fiction strategy game, and Schaeffer suggests that DeepMind devise a program that can learn to play different types of game from scratch. Such programs already compete annually at the International General Game Playing Competition, which is geared towards creating a more general type of AI. Schaeffer suspects that DeepMind would excel at the contest. "It's so obvious, that I'm positive they must be looking at it," he says.

DeepMind's founder and chief executive Demis Hassabis mentioned the possibility of training a version of AlphaGo using self-play alone, omitting the knowledge from human-expert games, at a conference last month. The firm created a program that learned to play less complex arcade games in this manner in 2015. Without a head start, AlphaGo would probably take much longer to learn, says Bengio — and might never beat the best human. But it's an important step, he says, because humans learn with such little guidance.

DeepMind, based in London, also plans to venture beyond games. In February the company founded DeepMind Health and launched a collaboration with the UK National Health Service: its algorithms could eventually be applied to clinical data to improve diagnoses or treatment plans. Such applications pose different challenges from games, says Oren Etzioni, chief executive of the non-profit Allen Institute for Artificial Intelligence in Seattle, Washington. "The universal thing about games is that you can collect an arbitrary amount of data," he says — and that the program is constantly getting feedback on what's a good or bad move by playing many games. But, in the messy real world, data — on rare diseases, say — might be scarcer, and even with common diseases, labelling the consequences of a decision as 'good' or 'bad' may not be straightforward.

Hassabis has said that DeepMind's algorithms could give smartphone personal assistants a deeper understanding of users' requests. And AI researchers see parallels between human dialogue and games: "Each person is making a play, and we have a sequence of turns, and each of us has an objective," says Bengio. But they also caution that language and human interaction involve a lot more uncertainty.

DeepMind is fuelled by a "very powerful cocktail" of the freedoms usually reserved for academic researchers, and by the vast staff and computing resources that come with being a Google-backed firm, says Joelle Pineau, a computer scientist at McGill University in Montreal. Its achievement with Go has prompted speculation about when an AI will have a versatile, general intelligence. "People's minds race forward and say, if it can beat a world champion it can do anything," says Etzioni. But deep reinforcement learning remains applicable only in certain domains, he says: "We are a long, long way from general artificial intelligence."

DeepMind's approach is not the only way to push the boundaries of AI. Gary Marcus, a neuroscientist at New York University in New York City, has co-founded a start-up company, Geometric Intelligence, to explore learning techniques that extrapolate from a small number of examples, inspired by how children learn. In its short life, AlphaGo probably played hundreds of millions of games — many more than Lee, who still won one of the five games against AlphaGo. "It's impressive that a human can use a much smaller quantity of data to pick up a pattern," says Marcus. "Probably, humans are much faster learners than computers." ■

MORE ONLINE

THE GO FILES

GOOGLE DEEPMIND

Updates from a battle of man versus machine over the Go board go.nature.com/j2Go3w

MORE NEWS

- Horse-sized dinosaur sheds light on *T. rex*'s origins go.nature.com/ihlkx2
- Peculiar pattern found in 'random' prime numbers go.nature.com/hz11qo
- Genetic study links 'good' cholesterol mutation to heart disease go.nature.com/mjils9

NATURE PODCAST

Retrieving lost memories; nailing China's emissions; and is Alzheimer's transmissible? nature.com/nature/podcast

HUMAN EVOLUTION

Ancient DNA pinpoints dawn of Neanderthals

Sequencing of 430,000-year-old DNA pushes back species' divergence from humans.

BY EWEN CALLAWAY

Matthias Meyer has just published the results of what may be the world's most wasteful genome-sequencing project. In decoding just 0.1% of the genome of the oldest DNA ever recovered from an ancient human, the molecular biologist at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, threw out enough raw data to map the modern human genome dozens of times over.

But the excess was necessary, because the DNA in the 430,000-year-old bones was degraded and contaminated. Meyer's feat of recovery has revealed that the remains, from a cavern in northern Spain, represent early Neanderthals — and has pushed back estimates of the time at which the ancient predecessors of humans must have split from those of Neanderthals (M. Meyer *et al. Nature* http://dx.doi.org/10.1038/nature17405; 2016).

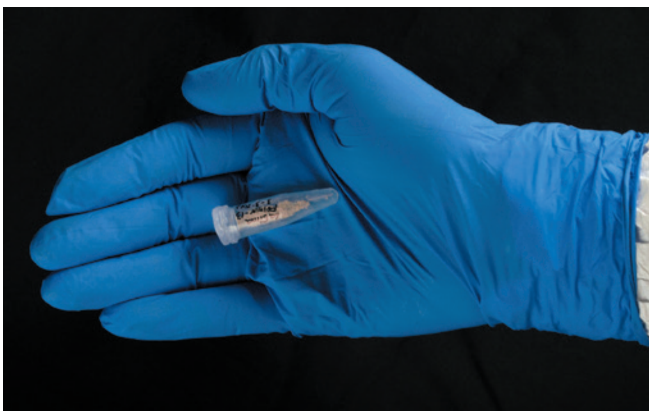
"Starting such a thing is already very ambitious, and managing it is even more impressive," says Ludovic Orlando, an ancient-DNA researcher at the Natural History Museum of Denmark in Copenhagen. "We are really reaching the limits of what is possible."

The analysis addresses confusion over which species the remains belong to. A report published in 2013 sequenced a femur's mitochondrial genome — which is made up of DNA from the cell's energy-producing structures that is more abundant in cells than is nuclear DNA. It suggested that at least one individual identified from the remains was more closely related to a group called Denisovans — known from remains found thousands of kilometres away in Siberia — than it was to European Neanderthals (M. Meyer *et al. Nature* **505,** 403–406; 2014).

"It's wonderful news to have mitochondrial and nuclear DNA from something that is 430,000 years old. It's like science fiction. It's an amazing opportunity," says Maria Martinón-Torres, a palaeoanthropologist at University College London.

The remains are known as the Sima hominins because they were found in Sima de los Huesos (Spanish for 'pit of bones'), a 13-metre-deep shaft in Spain's Atapuerca mountains. Few ancient sites are as important or intriguing as Sima, which holds the remains of at least 28 individuals, along with those of dozens of cave bears and other animals. The hominins might have plummeted to their death, but some researchers think they were deliberately buried there.

The Sima hominin skulls have the beginnings of a prominent brow ridge, as well as other traits typical of Neanderthals. But other features, and uncertainties around their age — some studies put them at 600,000 years old, others closer to 400,000 — convinced many researchers that they might instead belong to an older species known as *Homo heidelbergensis*.



Bone powder from the femur of a 430,000-year-old 'Sima' skeleton.

JAVIER TRUEBA, MADRID SCIENTIFIC FILMS

Confusion peaked when Meyer, his colleague Svante Pääbo and their team revealed the mitochondrial connection to the Denisovans. But they hoped that retrieving the skeletons' nuclear DNA — which represents many more lines of ancestry than does mitochondrial DNA, which is inherited solely from the maternal line — would clear things up.

NUCLEAR RECOVERY

Meyer's team managed to glean nuclear and mitochondrial DNA from five Sima samples, probably representing different individuals. A key factor in their success, says Meyer, was that since 2006, archaeologists had carefully refrigerated teeth and shoulder-blade tissue from the pit to preserve the ancient DNA — awaiting advanced molecular-analysis techniques.

The nuclear DNA, Meyer's team reports in *Nature* on 14 March, shows that the Sima hominins are in fact early Neanderthals. And its age suggests that the early predecessors of humans diverged from those of Neanderthals between 550,000 and 765,000 years ago — too far back for the common ancestors of both to have been *Homo heidelbergensis*, as some had posited.

Researchers should now be looking for a population that lived around 700,000 to 900,000 years ago, says Martinón-Torres. She thinks that *Homo antecessor*, known from 900,000-year-old remains from Spain, is the strongest candidate for the common ancestor, if such specimens can be found in Africa or the Middle East.

The team's latest mitochondrial sequences, meanwhile, again confirm the puzzling link between the Sima hominins and the Denisovans. Meyer suggests that the ancestors of the two groups carried mitochondrial DNA that is reflected in both — but which is not present in later Neanderthals.

This elimination could have happened by chance, but Meyer now favours the hypothesis that an as yet unknown species from Africa migrated to Eurasia and bred with Neanderthals, replacing the mitochondrial DNA lineages. (Supporting this idea, stone-tool technologies spread from Africa to Eurasia around half a million years ago, and again 250,000 years ago).

It is hard to rule out these or other ideas without new data, says Meyer. The full or nearly full genome of a Sima hominin, or genetic data from other early Neanderthals, would be necessary.

"It's fascinating and keeps us all on our toes trying to make sense of it all," says Chris Stringer, a palaeoanthropologist at the Natural History Museum in London. Stringer says that the recovery of such old nuclear DNA gives him hope that researchers will be able to analyse ancient DNA that stretches even further back in time. "Instead of just being stuck with trying to resolve the last 100,000 years," he says, "we can really start to put some dates from DNA further down the human tree." ■

MATHEMATICS

Fermat proof earns theorist Abel Prize

Andrew Wiles rewarded for cracking historic puzzle.

BY DAVIDE CASTELVECCHI

British number theorist Andrew Wiles has received the 2016 Abel Prize for his solution to Fermat's last theorem — a problem that stumped some of the world's greatest minds for three and a half centuries. The Norwegian Academy of Science and Letters announced the award — considered by some to be the 'Nobel of mathematics' — on 15 March.

Wiles, who is 62 and now at the University of Oxford, UK, will receive 6 million kroner (US$700,000) for his 1994 proof of the theorem, which states that there cannot be any positive whole numbers x, y and z such that $x^n + y^n = z^n$, if n is greater than 2. The award came as a "total surprise", he told *Nature*.

That Wiles solved a problem considered too hard by so many — and yet a problem that is relatively simple to state — has made him arguably "the most celebrated mathematician of the twentieth century", says Martin Bridson, director of Oxford's Mathematical Institute, which is housed in a building named after Wiles. Although his achievement is now two decades old, he continues to inspire young minds, something that is apparent when schoolchildren show up at his public lectures. "They treat him like a rock star," Bridson says. "They line up to have their photos taken with him."

"As human beings, we succeed by trial and error. It's the people who overcome the setbacks who succeed."

Wiles's story has become a classic tale of tenacity and resilience. While a faculty member at Princeton University in New Jersey in the 1980s, he embarked on a solitary, seven-year quest to solve the problem, working in his attic without telling anyone except his wife. He went on to make the historic announcement of his achievement at a conference in his hometown of Cambridge, UK, in June 1993, only to hear from a colleague two months later that his proof contained a serious mistake.

With another frantic year of work — and the help of one of his former students, Richard Taylor, who is now at the Institute for Advanced Study in Princeton — he was able to patch up the proof. When the resulting two papers were published in 1995, they made up an entire issue of the *Annals of Mathematics*[1,2].

But after Wiles's original claim had already made front-page news around the world, the pressure on the shy mathematician to save his work almost incapacitated him. "Doing mathematics in that kind of overexposed way is certainly not my style, and I have no wish to repeat it," he said in a BBC documentary in 1996, still visibly shaken by the experience.

"Unfortunately, as human beings, we succeed by trial and error," he told *Nature* after hearing about his win. "It's the people who overcome the setbacks who succeed." ■

1. Wiles, A. *Ann. Math.* **141,** 443–551 (1995).
2. Taylor, R. & Wiles, A. *Ann. Math.* **141,** 553–572 (1995).

SPACE

Mars launch puts Russia–Europe team to the test

Joint ExoMars mission launches lander and orbiter, with a rover planned for 2018.

BY ELIZABETH GIBNEY

Neither Europe nor Russia has ever successfully operated a mission on Mars's surface. Now the European Space Agency (ESA) and its Russian counterpart Roscosmos hope to mark a first for both organizations with the joint ExoMars 2016 mission, which launched from the Baikonur Cosmodrome in Kazakhstan on 14 March.

ExoMars 2016 consists of a lander that will study the planet's dust storms and an orbiter that will analyse its atmosphere, including looking for methane. The orbiter will also act as a relay for a Mars rover, due to be launched in 2018. Each phase will test the growing collaboration between the two agencies, which have hinted at further joint missions, including uncrewed and crewed Moon trips.

STEPHANE CORVAJA/ESA

A Russian Proton rocket launches ExoMars 2016 from Kazakhstan.

ESA designed the orbiter and lander projects, but a Russian Proton rocket launched them, and they carry Russian instruments. "The launch is crucial because it's symbolic," says Oleg Korablev of the Space Research Institute, Moscow, who is principal investigator for the Atmospheric Chemistry Suite on the orbiter. "It's psychologically very important."

ESA project scientist Jorge Vago adds: "Hopefully this will cement a way of doing things that becomes the modus operandi for when we do missions together." Based at the European Space Research and Technology Centre in Noordwijk, the Netherlands, Vago also works on the ExoMars 2018 mission.

ESA approved the ExoMars concept in 2005; a subsequent merry-go-round of collaborators eventually resulted in the Europe–Russia collaboration and ExoMars's unusual two-stage format (see 'Mission merry-go-round').

The ExoMars 2016 craft — the heaviest Mars mission ever launched, at 4,332 kilograms — is now bound for the red planet. A favourable alignment of Earth and Mars means that the craft should reach orbit in 7 months; the orbiter and landing module, Schiaparelli, will separate before reaching the Martian atmosphere.

The landing won't involve anything as complex as NASA's sky crane, which delivered the Curiosity rover to Mars in 2012. But it is ambitious, says Vago, and designed to show that Europe can make a controlled landing on Mars. Heeding lessons from Beagle 2 (Britain's failed 2003 Mars lander that was operated by ESA), the module will use drag from the Martian atmosphere to brake, then open a parachute and finally fire its thrusters. During the last 2 metres, Schiaparelli will deploy a honeycomb-like crash pad.

The first lander to set down during dust-storm season, Schiaparelli will monitor pressure and temperature and image the approaching landing site as it descends. On the ground, the conical lander has just 2–4 days of battery power to perform experiments.

Its tiny meteorological station DREAMS (Dust Characterisation, Risk Assessment, and Environment Analyser on the Martian Surface) will measure pressure, humidity, temperature, wind speed and direction. This represents a unique chance to study dust circulation and hopefully unravel the mystery of why some storms on Mars go planet-wide, says Francesca Esposito at the INAF Astronomical Observatory of Capdiomonte in Naples, Italy, and the principal investigator for DREAMS. The lander will also be the first to examine the planet's electric field. Those data will feed into Martian climate models and could allow scientists to better predict future disturbances to communications on the planet, she says.

BIOLOGICAL STUDIES

The higher-profile science — including investigating hints of Martian biology — will take place in the sky on board ExoMars' Trace Gas Orbiter (TGO). That phase will begin at the end of 2017, once the craft has manoeuvred into a circular, 400-kilometre-high orbit.

While studying Mars's atmosphere, the TGO's major task will be to follow up on evidence that the red planet contains methane, which has been associated with active geological processes, as well as biological ones. "Is there a seasonality to the methane, or are the concentrations associated with particular types of terrain, for instance?" asks John Bridges, a planetary scientist at the University of Leicester, UK, who works on the TGO's stereo camera. The camera will use 3D images to chart geological features; a hydrogen detector will map the planet's subsurface water.

The orbiter will also serve as a communications platform, including for the ExoMars rover, which will break new ground — literally. The rover will drill up to 2 metres into the surface, where organic matter, which can be destroyed by surface radiation, may lie preserved.

The rover project will require the Russian and European teams to work together to an extent unprecedented for ESA, says Vago. In the 2016 mission, their responsibilities are relatively separate, but in the 2018 mission, he says, "there is no clean line", and each design tweak ripples through the work of both teams. It is a new experience for Russia too, says Korablev, even though the country has long contributed scientific instruments to foreign space missions. "There are many problems, but there are always problems on national projects too," he says.

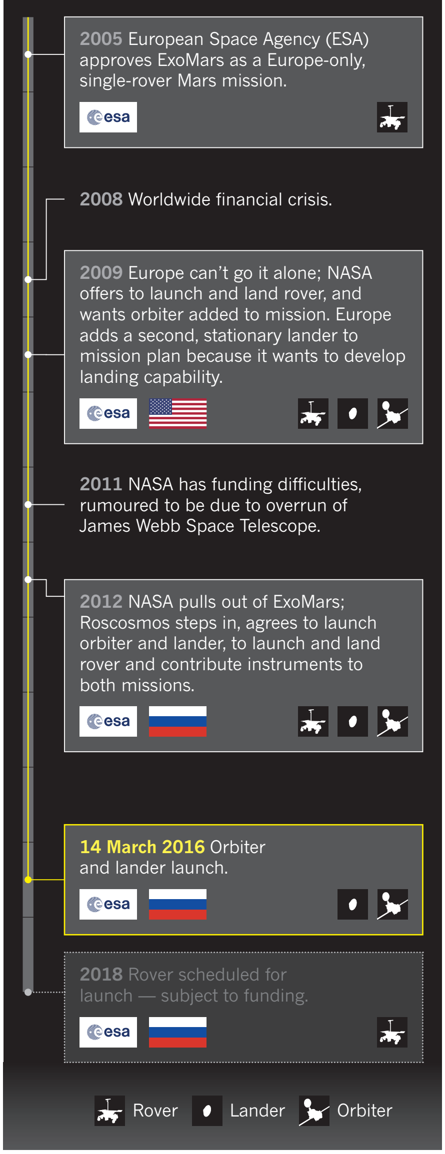
This complexity, coupled with late-running instruments, delays in testing and a lack of cash, means that the 2018 rover mission could be delayed until 2020, says Vago. ESA's

director-general, Johann-Dietrich Wörner, said in January that the 2018 mission needed more funding to meet its launch target, and the agency is expected to ask member states for the missing few hundred million euros at a meeting in December. Success with ExoMars 2016 could help to persuade European leaders to contribute. Bridges says that most scientists will accept a delay as long as it means that all the instruments are on the craft and working.

Korablev's involvement with Mars missions has been an emotional roller coaster. He spent 10 years working on Russia's Mars 96 orbiter, which failed to leave near-Earth orbit. He was also involved in a sample-return mission to the Martian moon Phobos, which ran into problems, eventually crashing in the Pacific Ocean in 2012. "We put a great effort into ExoMars," he says. "I almost don't dare to say any words." ■

MISSION MERRY-GO-ROUND

The countries involved in the ExoMars mission, as well as the mission components, have changed multiple times since its conception.



PAUL JACKMAN/NATURE

PLANETARY SCIENCE

NASA Mars woes could delay missions

Decision to defer InSight launch will cost US$150 million.

BY DEVIN POWELL

Cracks in an instrument designed to detect earthquakes on Mars will add roughly US$150 million to the price tag of InSight, NASA's next mission to the red planet. But the agency said on 9 March that it still intends to fly the spacecraft, raising questions about how the unexpected expense will affect other planetary missions in development.

Although InSight's launch — originally scheduled for this month — is now slated for May 2018, it is not clear whether the spacecraft's faulty seismometer will be ready in time. InSight seeks to investigate Mars's interior by measuring seismic activity, as well as the heat that is escaping from the planet and the movement of its surface.

"We're really grateful that NASA has recognized the value of science we're going to do and agreed to give us a chance to try it again," says InSight's principal investigator, Bruce Banerdt, who works at NASA's Jet Propulsion Laboratory (JPL) in Pasadena, California.

The spacecraft was developed as part of NASA's Discovery Program, which funds small, quick-turnaround missions whose costs are capped at $450 million. Five proposed missions are currently vying for a chance to launch in the early 2020s. These include a trip to Jupiter's Trojan asteroids, two missions to Venus and a camera that would detect near-Earth objects.

"NASA has been trying to choose two missions out of this round instead of one, and the community's concern is that the likelihood of that happening might be falling," says Linda Elkins-Tanton, a planetary scientist at Arizona State University in Tempe and principal investigator of a proposed mission to the asteroid Psyche.

Jim Green, director of NASA's planetary-science division in Washington DC, says that budget details and consequences to other planetary missions will be worked out by August. "Our ability to select at least one Discovery mission in December is expected to be unaffected," he says.

Then there is the matter of whether InSight's troubled seismometer, which was developed by a global collaboration led by the Paris Institute of Earth Physics, can be repaired.

Banerdt says that Sodern, the French company subcontracted to build a vacuum container to enclose the seismometer's sensors, did not detect any problems with connectors that are supposed to seal wires leading out of the vacuum housing. Only when the instrument was tested in frigid, Mars-like temperatures in December did cracks in those seals become apparent. The project team tried to patch the problem, but persistent leaks remained.

"It's very frustrating," says Banerdt. "I've been working on getting this kind of mission for more than 25 years, and everything else on the project was going really well."

FUTURE FIX

NASA has asked the JPL to craft a new, hardier vacuum chamber. The agency's French collaborators will test the chamber at their own expense. "Personally, I am relieved to know that JPL will be taking responsibility for the vacuum chamber," says Lisa Pratt, a biogeochemist at Indiana University Bloomington.

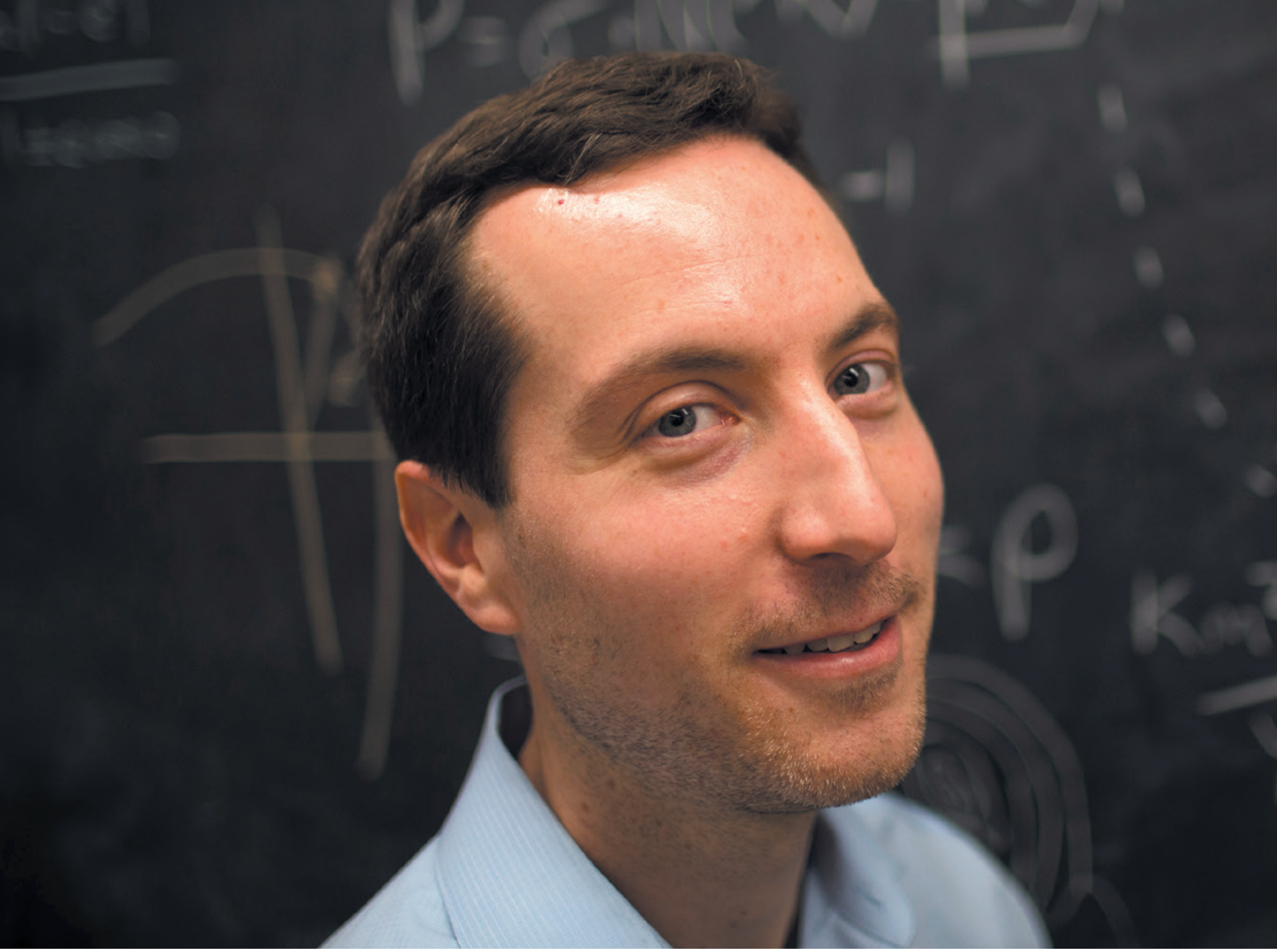
The Mars InSight team is now re-running landing simulations and recalculating orbits to account for the updated launch date.

"We're really grateful that NASA has recognized the value of science we're going to do."

Some wonder whether the mistake could cause NASA to tighten the reins on future projects. The most recent call for Discovery mission proposals — made before the problem with InSight occurred — mandated that no more than one-third of instrument costs be spent on foreign sources.

"The word on the street is that NASA's a little more wary of collaborating with groups that they don't know so well or don't control directly," says Elkins-Tanton.

But Green argues that any nation trying to build a new instrument could have made this mistake. "This is the first time this type of instrument has been built to withstand harsh environmental conditions on another planet," he says. ■



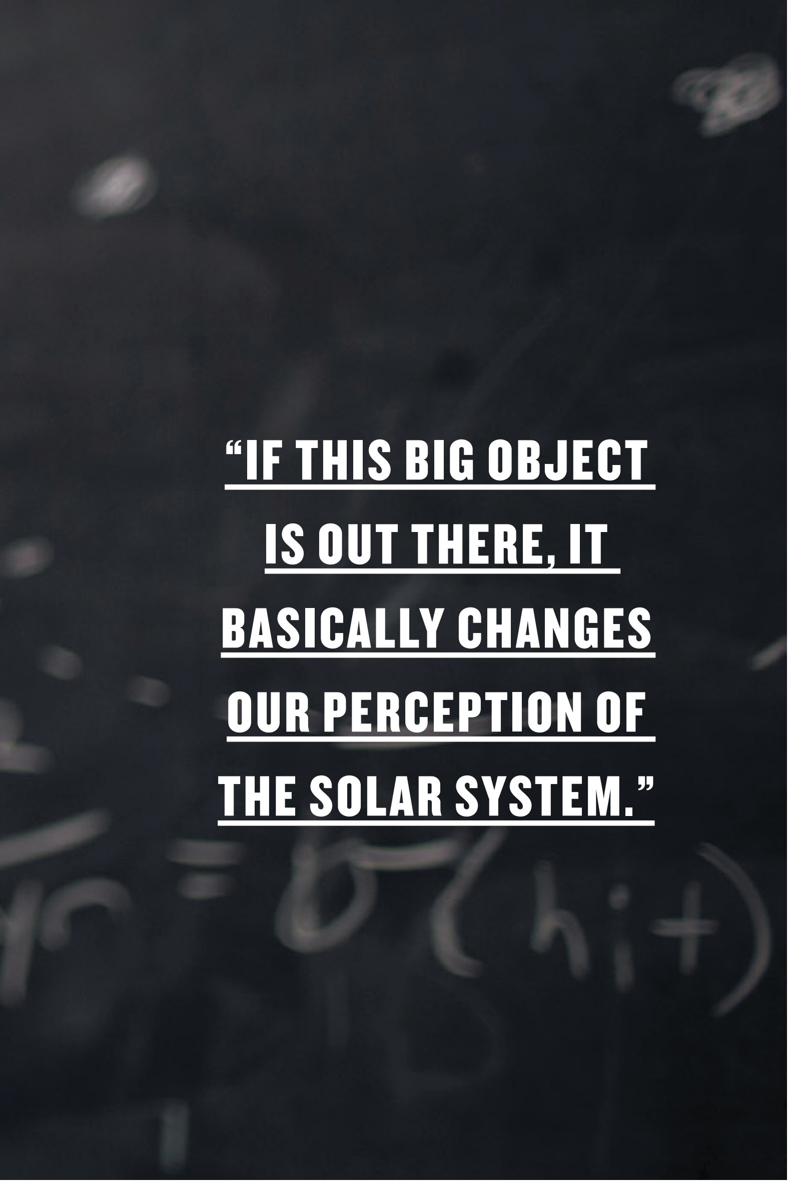
ON THE HUNT FOR A MYSTERY PLANET

SCIENTISTS ARE SEARCHING FOR AN UNSEEN WORLD AT THE FRINGES OF THE SOLAR SYSTEM.

BY ALEXANDRA WITZE

Astronomer Scott Sheppard runs through his checklist as he settles in for a long night of skygazing at the Subaru telescope atop Mauna Kea, Hawaii. The air above the summit: clear. The telescope: working smoothly. His 3-terabyte hard drive: emptied and ready to accept a flood of fresh data in the hours to come.

On a wall in the observing room, three clocks track the hours in Hawaii, Tokyo and Coordinated Universal Time. Screens display every tic of the weather above the summit: wind direction, temperature and the dreaded humidity levels that could end this November night of observing if they were to rise. But for now, conditions are nearly perfect, especially when it comes to a characteristic known as seeing — a measure of how stable the stars above look. "Seeing is point-five-five," says David Tholen, an astronomer at the University of Hawaii in Manoa. "It doesn't get much



Astronomer Scott Sheppard is searching systematically for distant objects — one of which could be the elusive Planet X.

CHRIS MADDALONI/*NATURE*

better than that," replies the third member of this team, Chad Trujillo of the Gemini Observatory in Hilo, Hawaii. Sheppard, the lone mainlander of the group, works at the Carnegie Institution of Science in Washington DC. With the weather looking promising, he pulls out his logbook and begins to outline plans for the next ten hours.

Between twilight and dawn, he will methodically direct Subaru's enormous 8.2-metre mirror — one of the largest in the world — to stare deeply at one patch of the sky, then another and another. Several hours later, he will look at the same areas for a second time, and after that, a third. By comparing the staggered images, the researchers can hunt for objects that move ever so slightly over the course of a few hours. These would be distant worlds beyond Pluto, in the most extreme reaches of the Solar System. This is the realm of the long-sought Planet X.

Sheppard already has some idea of where to look. On his list of goals for the night, he wants to capture fresh images of an object that he first spotted one month earlier. At 9:20 p.m., and again at 10:46 p.m., he aims Subaru at a region near the constellation Aries, where he thinks this object will be. The exposures come off the telescope, and Tholen begins to process them. After a few minutes, he beckons Sheppard over. On his grey screen is a light-coloured dot that jumps against the field of background stars when Tholen toggles between the two images. "There it is," he says. "You got it."

"Right where it should be," Sheppard says. It is the same object that he saw previously, and it earns an entry in his logbook: in an image on computer chip number 104, in field number 776, they have spotted an object 90 astronomical units (AU) away, or 90 times the Earth–Sun distance. It's not clear yet how big the object is or whether it is scientifically important — but it is among the most distant worlds ever seen in the Solar System.

Astronomers have discovered more than 2,000 exoplanets around other stars, mostly through indirect methods that detect changes in the distant star. Yet the farthest reaches of our Solar System remain largely unexplored: the indirect techniques won't work in our own stellar neighbourhood, and objects in the distant suburbs of the Sun are too faint to be glimpsed by anything but the world's most powerful telescopes. Sheppard and Trujillo have been racing to find the frigid worlds that are thought to populate this distant zone. At stake is what could be the last great discovery in the Solar System — a planet bigger than Earth that may swing around the Sun far past Pluto.

Suggestions of a Planet X have circulated for more than a century, but those hypotheses have always fallen apart after closer scrutiny. In 2014, Trujillo and Sheppard revived the concept of a putative Planet X, on the basis of the orbits of some extremely distant objects[1]. In January, the idea got a boost when two astronomers from the California Institute of Technology (Caltech) in Pasadena calculated more precisely where in the Solar System it might lie[2]. They dubbed it 'Planet Nine': a backhanded reference to the demotion of Pluto from planet to dwarf planet in 2006.

The hunt is now on to find Planet Nine, or any other unseen super-Earths that may lurk out there on other orbits. The quest is likely to reveal fundamental insights into how the Solar System formed 4.6 billion years ago, and how it has evolved since then. "If this big object is out there, it basically changes our perception of the Solar System," says Sheppard. "It's true discovery at its core."

UNKNOWN ZONE

When Sheppard and Trujillo began to chase distant worlds, they were hoping to follow in the footsteps of other legendary astronomers. In 1846, Johann Galle of Germany first spotted Neptune, the eighth planet, at a distance of about 30 AU, right where it was expected to be according to calculations of how it would gravitationally perturb Uranus. In 1930, US astronomer Clyde Tombaugh found Pluto, orbiting at a distance of around 40 AU. And in 1992, astronomers David Jewitt, then at the University of Hawaii, and Jane Luu, then at the University of California, Berkeley, discovered an even more distant object, beginning the exploration of a region of space called the Kuiper belt[3].

Since then, astronomers have found thousands of Kuiper belt objects: small icy worlds similar to Pluto that range in distance from about 30 AU to 50 AU from the Sun. Beyond that lies the equivalent of 'here be dragons' on old maps. Scientists sometimes call it the outermost Kuiper belt or the innermost Oort cloud — the next region of the Solar System, which is thought to extend to at least 100,000 AU. "There's a whole chunk of the Solar System we don't fully understand," says Meg Schwamb, a planetary astronomer at the Academia Sinica in Taipei. "It's one of the last unexplored territories."

Which is why Sheppard and Trujillo are out looking. They met as graduate students at the University of Hawaii, where both had Jewitt as their adviser. They worked together to hunt Kuiper belt objects, and then started a systematic survey to search for still more-distant worlds. They are the only team routinely looking for the most-extreme objects. "The population could be huge," says Trujillo. "That's why we're doing the search."

By 2012, the two were using the biggest light buckets they could get their hands on, with wide-field cameras that would let them view as much of the sky as possible. At the Dark Energy Camera, atop a 4-metre telescope in Chile, they got a hit almost immediately. On their first night of observing, they spotted an object that was moving so slowly that it had to be very distant. Thrilled, they watched it move during the course of a year, which provided enough data to calculate its orbit.

They found that the closest it ever gets to the Sun — a measure known as its perihelion — is 80 AU, beyond the bulk of the Kuiper belt. This made it the object with the farthest-known perihelion, just beating the

dwarf planet Sedna, whose perihelion is 76 AU.

The discovery of this object, called 2012 VP_{113}, led to a *Nature* paper[1] — and a lot more observing time on big telescopes. In 2014, Sheppard and Trujillo spent their first nights at Subaru, a facility run by the National Astronomical Observatory of Japan that carries a huge camera called the Hyper Suprime-Cam. The combination of a big telescope and a wide-field camera makes Subaru the world's best place to scan large sections of the sky for faint objects.

Many scientists work with Subaru remotely: they stay at sea level in Hilo, and use videoconferencing to communicate with the telescope's operators. That approach saves researchers from making the 2-hour-long journey to the summit of Mauna Kea at 4,200 metres, where the atmosphere has 40% less oxygen and causes many people to experience dizziness, headaches or sometimes more-serious medical problems.

But Sheppard likes to be actively involved in directing observations, so he always makes the trip. As the hours tick by during the night, he stays alert, never once clipping an oxygen sensor onto his finger to see how he is coping with the altitude. His logbook fills up with notations: field number, chip number, exposure time. He reorders targets on the fly, rearranging what he is looking at to improve the time gap between the fields.

Subaru's huge mirror gazes at the sky, gathering photons for him. Exposure times count down in big green numbers on a computer monitor. When an exposure finishes, an alert dings like a cuckoo clock, and Sheppard hovers over the shoulder of the telescope operator to tell him where to point the camera next.

Each good observing night fills up Sheppard's Macbook with data. To identify potential distant worlds, the researchers use a programme that Trujillo wrote to pick out objects that move between different frames of the same star field. But because the programme flags a number of false positives, each field must also be reviewed manually. Sheppard goes through every exposure, eyeballing the faint dots that the programme circled in orange to decide whether they represent a distant Solar System object or something else — an asteroid or a cosmic-ray blip, perhaps.

Frame by frame, Sheppard zips through thousands of exposures as if he's playing a video game. "It's exciting to go through," he says. "Every image — you never know what you're going to get. It could be the image with the super-Earth in it."

"EVERY IMAGE — YOU NEVER KNOW WHAT YOU'RE GOING TO GET. IT COULD BE THE IMAGE WITH THE SUPER-EARTH IN IT."

The key is how slowly the objects move. Asteroids are relatively close to Earth, and their position in the sky can shift by 30 arcseconds, or 0.008 degrees, each hour. Kuiper belt objects, which are much farther away, traverse about 3 arcseconds of sky each hour. Anything slower than that must be beyond the main Kuiper belt and is something that interests the search team.

Astronomers must observe an object multiple times over the course of a year to pin down its orbit and determine its perihelion. Just because an object is remote when it is discovered does not mean that it is scientifically important.

For instance, the object that Sheppard spotted at 90 AU in November may have been at its closest approach to the Sun. If so, that would make it a record-breaker, situated beyond Sedna and 2012 VP_{113}. Or the object might be travelling on a path that takes it much, much closer to the Sun, perhaps 40 AU. That would make it less exciting, because its perihelion distance would place it squarely within the main Kuiper belt — meaning that it is just another ordinary Kuiper belt object, as opposed to one of the extreme worlds.

The same is true for an object that the scientists found at 103 AU last November — the most distant ever observed. They will not know for many months whether that body stays in the outer Solar System, or whether it veers inward at its perihelion.

By far the most prized quarry out there is the hypothesized Planet Nine. In their 2014 *Nature* paper, Trujillo and Sheppard suggested — on the basis of the orbits of 2012 VP_{113} and Sedna — that an unseen super-Earth could lurk at roughly 250 AU. This January, Konstantin Batygin and Mike Brown of Caltech took these two bodies, along with four other distant Kuiper belt objects, and compared their orbits to narrow the calculations of where such a planet might lie.

All six objects share a common orbital property: when they pass closest to the Sun, they are travelling from north to south relative to the plane of the Solar System. If they had no relation to one another, they should not all share that orientation. A second line of argument is that the six objects are also physically clustered in space (see 'Out there'). "They all point in the same direction and are all tilted at the same angle," says Batygin. "That's odd."

He and Brown argue that an unseen Planet Nine must be shepherding them into those clusters. It would be between 5 and 10 times the mass of Earth, and travel as close as 200 AU to the Sun and as far away as 1,200 AU.

FINDING THE NINTH

Critics say that the argument rests on just a handful of weird Kuiper belt objects. "It's very small statistics," says David Nesvorný, a planetary scientist at the Southwest Research Institute in Boulder, Colorado, who nonetheless finds the concept intriguing. "It's as science should be — at the edge of believability."

Many astronomers are now running their own calculations to estimate the chances that Planet Nine exists in this particular orbit, and if not, where it might be. Samantha Lawler, of National Research Council–Herzberg in Victoria, Canada, is working with Nathan Kaib, of the University of Oklahoma in Norman, to explore how the presence of a super-Earth might affect the orbits of many Kuiper belt objects. Their preliminary results suggest that, if a Planet Nine were out there, it should have nudged the orbits of Kuiper belt objects in ways that do not reflect reality.

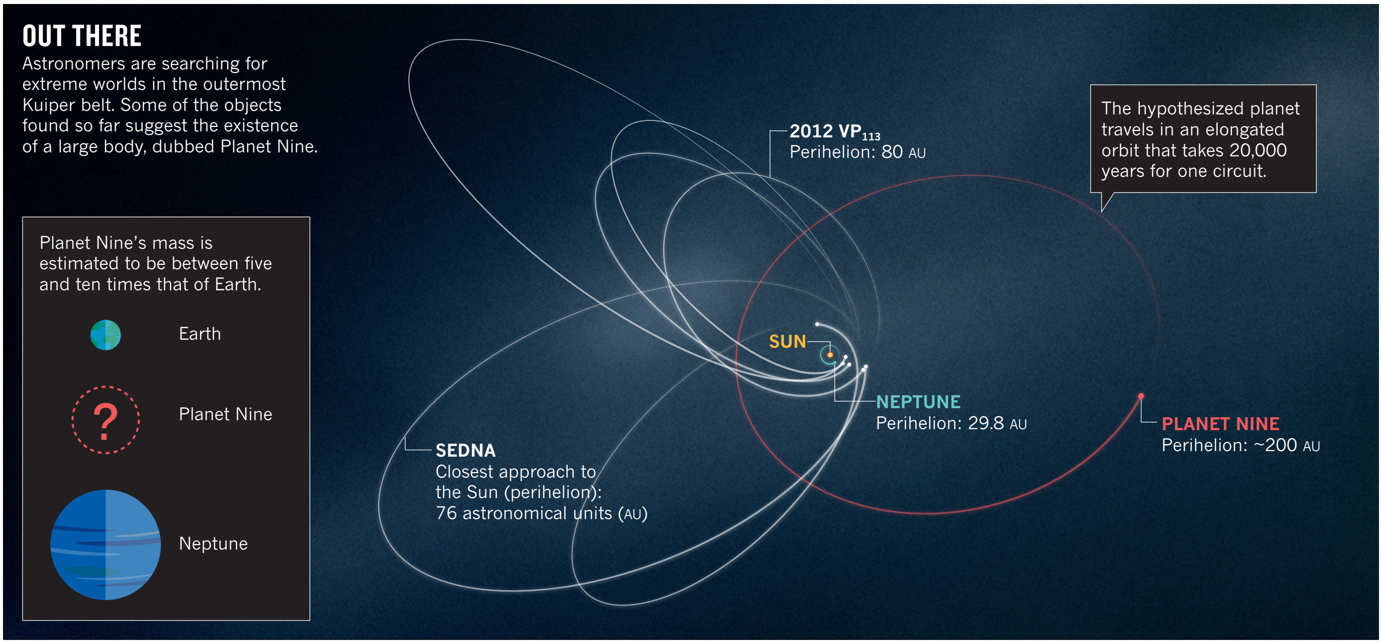
Planet Nine "is a cool idea, and it would be really neat if it was true", says Lawler. "But you have to be really careful."

Some answers may come from an ongoing project known as the Outer Solar System Origins Survey (OSSOS), run by a consortium of investigators. It is working to find and study all the observable Kuiper belt objects in a small patch of the sky in extraordinary detail — by following their orbits, classifying their colours and so on. That work has the potential to rule out the existence of Batygin and Brown's hypothesized Planet Nine — if OSSOS were to find a distant object in a region that should have been cleared out by the proposed planet.

Other astronomers have suggested alternative ways to hunt Planet Nine, such as looking at data from the Cassini spacecraft orbiting Saturn to see whether that planet's orbit is perturbed ever so slightly, or by using cosmological telescopes at the South Pole to detect a planet's faint radiation. As Sheppard and Trujillo continue their methodical survey of the sky, they are paying special attention to areas where Batygin and Brown say the planet could be. And the Caltech pair is chasing it as well, also using Subaru. "I'd be astonished if there isn't some kind of planet there," says Renu Malhotra, a theorist at the University of Arizona in Tucson. In a paper on the preprint server arXiv, she and her colleagues put forward a new analysis[4] of where a super-Earth might lurk, on a different orbit from Batygin and Brown's Planet Nine. Malhotra's team uses four extreme Kuiper belt objects to suggest that an unseen planet moves around the Sun every 17,000 years.

But even if a large planet is out there, it will take some luck to find it with existing technology. For one of the teams to spot the object, it

SOURCE: CALTECH/REF. 2



would have to be on the larger end of its estimated size range, or be very reflective or in a relatively close-in orbit. If the planet is too small, dark and far away, it may never be seen from Earth. "It's worse than looking for a needle in a haystack," says Malhotra. "It's like looking for the broken tip of a needle in a haystack."

THE STORY OF A PLANET

A more fundamental question is not whether Planet Nine exists, but what distant objects say about planetary evolution more generally. Discoveries such as Sedna and 2012 VP_{113} have forced a radical rethinking of the gravitational forces that shape the outer parts of the Solar System.

When astronomers first started to find Kuiper belt objects in the 1990s and recognized that Pluto was just another member of that clan, they began to paint a picture of this mysterious realm of space. The Kuiper belt seemed to extend neatly from about 30 AU to 50 AU, with most objects following stately orbits around the Sun. Those that were a bit odd — travelling off-kilter to the plane of the Solar System, or occasionally to greater distances — could be explained by gravitational interactions with Neptune.

Sedna and 2012 VP_{113} do not fit that simple model because they range too far from the Sun to have ever interacted much with Neptune. Theorists suddenly had to confront the question of how these objects reached their current orbits. All known planets in the Solar System, along with the Kuiper belt objects, are thought to have condensed from a disk of gas and dust that swirled around the newborn Sun 4.6 billion years ago. But Sedna and other objects beyond the main Kuiper belt probably weren't born where they are today, because there simply wasn't enough gas and dust available at those great distances to create sizeable worlds.

One idea is that they were tossed there by a gravitational battle with other protoplanets closer to the Sun during the first tens of millions of years of the Solar System's existence. A second theory holds that the gravity of a passing star tugged on the outer bits of the planet-forming disk, pulling nascent planets into elongated orbits, where they remain today.

If Planet Nine exists, it could complicate this picture even more. It would mean that the orbits of Sedna and 2012 VP_{113} were not fixed early on but are being actively shaped — even today — by the gravitational tugs of Planet Nine. That would require theorists to rewrite their ideas about how the Solar System's many worlds have interacted with one another over the past 4.6 billion years. "It's hard to anticipate what direction our imaginations will go," says Malhotra.

Understanding the distant Kuiper belt could also help astronomers to work out how our Solar System compares with planetary systems around other stars. Brown notes that one of the most common types of exoplanet is one missing from our Solar System, a world more massive than Earth but less massive than Neptune — that is, around the range of the hypothesized Planet Nine. "Maybe we can see what this most common type of planet might actually look like," he says.

For now, scientists' best shot at answering these questions is to find more extremely distant worlds. And that is why Sheppard and Trujillo keep plugging away in Chile and Hawaii, having covered less than 10% of the sky that they intend to survey.

Back on Mauna Kea, Sheppard pushes through the night of observing, clocking one field after another with no break. By 4:45 a.m., the atmosphere above the summit is turning a little more opaque, and he begins to shift the exposure times longer. Finally, at 5:25 a.m., he turns to the videoconferencing unit and calls to his colleague in Hilo. "Chad, are you there?" he asks. "All the fields are in." The skies above Subaru are beginning to brighten, although Sheppard does not get to enjoy the spectacular view of a Hawaiian sunrise because he does not step outside. He is busy tallying his 33 fields for the night. Any one of them could contain a new extreme Kuiper belt object — or even a Planet Nine.

It is after 7 a.m. when the observing team tumbles into two sports-utility vehicles and drives the steep, rocky road down from Mauna Kea's summit. Sheppard starts to flag only when he sits down for breakfast at the astronomers' dorm, 1,360 metres lower down on the mountain. He and Tholen gulp down their food and retreat to black-curtained dormitories to sleep until noon.

Sheppard, aged 40, has asked his doctor about eye strain and whether he will be able to keep looking at star fields forever. He and Trujillo have a self-appointed goal of finding ten inner-Oort-cloud objects, a number that they think will enable them to start testing ideas for how these objects formed and evolved.

That means many more long nights at the telescope. "If it turns into postage-stamp collecting, we'll stop," Sheppard says. "But right now, every new discovery is a huge difference-maker in trying to understand what's going on out there." ■

Alexandra Witze *writes for* Nature *from Boulder, Colorado. Not one to pull an all-nighter, she took a nap at Subaru.*

1. Trujillo, C. A. & Sheppard, S. S. *Nature* **507,** 471–474 (2014).
2. Batygin, K. & Brown, M. E. *Astron. J.* **151,** 22 (2016).
3. Jewitt, D. & Luu, J. *Nature* **362,** 730–732 (1993).
4. Malhotra, R., Volk, K. & Wang, X. Preprint at http://arxiv.org/abs/1603.02196v2 (2016).

THE RED-HOT DEBATE ABOUT TRANSMISSIBLE ALZHEIMER'S

A controversial study has suggested that the neurodegenerative disease might be transferred from one person to another. Now scientists are racing to find out whether that is true.

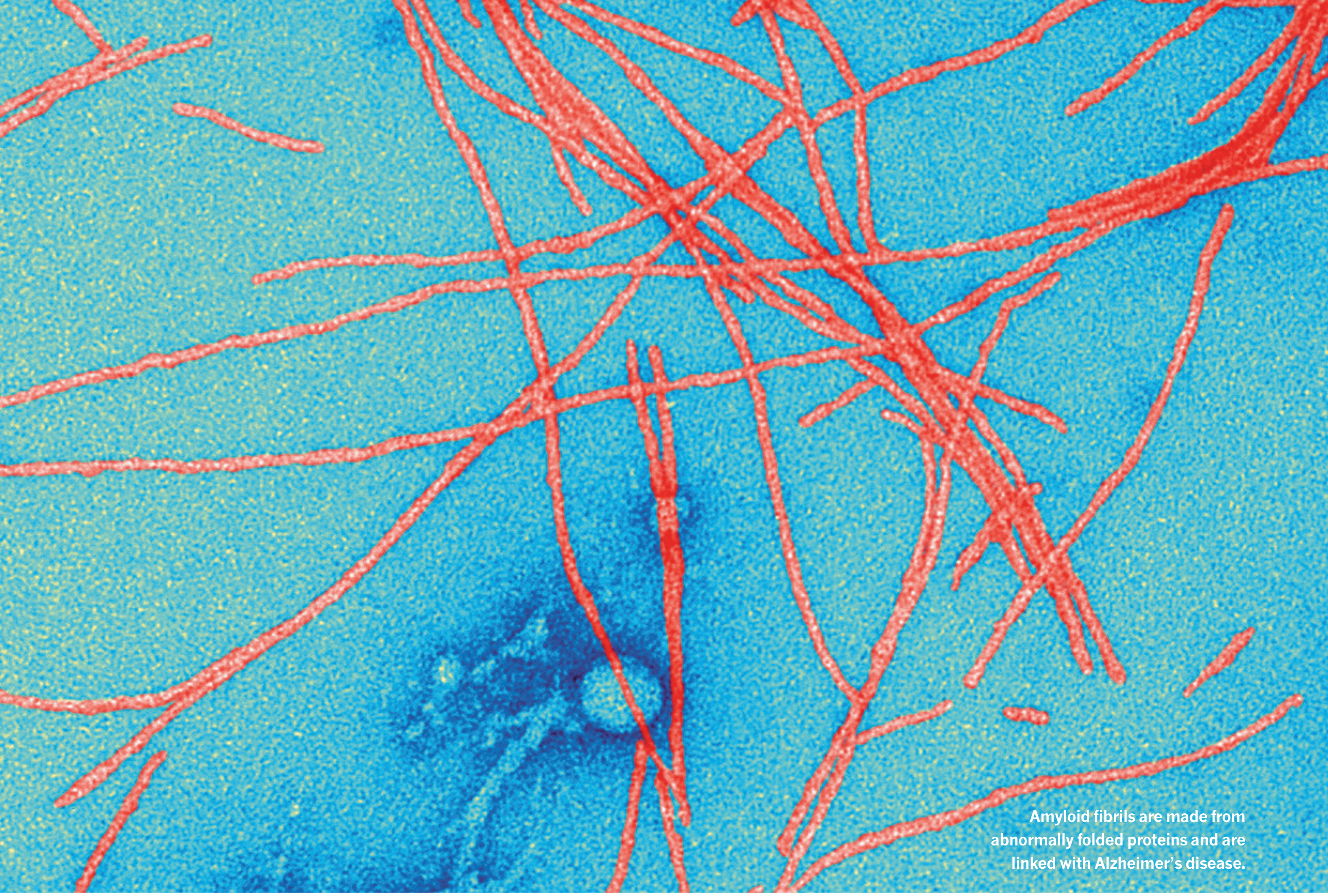
BY ALISON ABBOTT

In the 25 years that John Collinge has studied neurology, he has seen hundreds of human brains. But the ones he was looking at under the microscope in January 2015 were like nothing he had seen before.

He and his team of pathologists were examining the autopsied brains of four people who had once received injections of growth hormone derived from human cadavers. It turned out that some of the preparations were contaminated with a misfolded protein — a prion — that causes a rare and deadly condition called Creutzfeldt–Jakob disease (CJD), and all four had died in their 40s or 50s as a result. But for Collinge, the reason that these brains looked extraordinary was not the damage wrought by prion disease; it was that they were scarred in another way. "It was very clear that something was there beyond what you'd expect," he says. The brains were spotted with the whitish plaques typical of people with Alzheimer's disease. They looked, in other words, like young people with an old person's disease.

For Collinge, this led to a worrying conclusion: that the plaques might have been transmitted, alongside the prions, in the injections of growth hormone — the first evidence that Alzheimer's could be transmitted from one person to another. If true, that could have far-reaching implications: the possibility that 'seeds' of the amyloid-β protein involved in Alzheimer's could be transferred during other procedures in which fluid or tissues from one person are introduced into another, such as blood transfusions, organ transplants and other common medical procedures.

Collinge felt a duty to inform the public quickly. And that's what he did, publishing the study in *Nature* in September[1], to headlines around the world. "Can you CATCH Alzheimer's?" asked Britain's *Daily Mail*, about the "potentially explosive new study". Collinge has been careful to temper the alarm. "Our study does not mean that Alzheimer's is actually contagious," he stresses. Carers won't catch it on the job, nor family members, however close. "But it raises concern that some medical procedures could be inadvertently transferring amyloid-β seeds."



Amyloid fibrils are made from abnormally folded proteins and are linked with Alzheimer's disease.

SCIENCE SOURCE/SPL

Since then, the headlines have died away, but the academic work and discussion have taken off. Could seeds of amyloid-β proteins really be transmitted and, if so, are they harmless or do they cause disease? And could seeds of other related diseases involving misfolded proteins be transmitted in a similar way? In the past decade or so, evidence has been mounting for a controversial theory that rogue proteins, known collectively as amyloids and associated with diverse neurodegenerative diseases — from Alzheimer's to Parkinson's and Huntington's — might share some properties of prions, including their transmissibility. Collinge's data bolstered that theory.

Urgent though these questions are, it could take years to find answers. The paper by Collinge and his colleagues has sparked a worldwide hunt for similar amyloid pathology in autopsied brains, and a small study[2] published this January revealed a handful of related cases. Researchers are also trying to work out what the putative amyloid seeds look like, and whether different 'strains' of amyloids exist that are particularly damaging.

Some researchers say that it is much too early to be alarmed. They point out that the number of patients in Collinge's study was tiny, that none had displayed symptoms of Alzheimer's disease before their death and that another toxic protein called tau also seems to be required to cause the condition. "We have to remember that there is no conclusive evidence that seeds of amyloids can transmit actual disease or that amyloids spread in the brain in a prion-like way," says Pierluigi Nicotera, scientific director of the German Centre for Neurodegenerative Diseases in Bonn. "There may be other biological explanations."

Right now, there are few solid answers, but plenty of concerns. The sceptics worry that they might one day find themselves working under tight biosecurity regulations to handle proteins that they view as relatively innocuous. Others feel that the dangers may have been underestimated, and that scientists have a duty to investigate this as quickly as they can. "In my opinion, all amyloids should be considered dangerous until proven safe," says prion and amyloid researcher Adriano Aguzzi at the University Hospital Zurich in Switzerland.

DANGEROUS FOLDS

A few decades ago, it was almost inconceivable that a protein, which has no genetic material or any other obvious way to self-replicate, could cause infectious disease. But that changed in 1982, when Stanley Prusiner, now at the University of California, San Francisco, introduced evidence for disease-causing prions, coining the term from the words 'proteinacious' and 'infectious'[3]. Prusiner showed that prion proteins (PrP) exist in a normal cellular form, and in a misfolded infectious form. The misfolded form causes the normal protein to also misfold, creating a cascade that overwhelms and kills cells[4]. It cause animal brains to turn into a spongy mess in scrapie, a disease of sheep, and in bovine spongiform encephalopathy (BSE or 'mad cow disease'), as well as in human prion diseases such as CJD.

Prusiner and others also investigated how prions could spread. They showed that injecting brain extracts containing infectious prions into healthy animals seeds disease[4]. These prions can be so aggressive that in some cases, simply eating infected brains is sufficient to transmit disease. For example, many cases of variant CJD (vCJD) are now thought to have arisen in the United Kingdom in the 1990s after people ate meat from cattle that were infected with BSE.

➲ NATURE.COM
Listen to more on the transmissible Alzheimer's debate:
go.nature.com/jvfggn

Since then, scientists have come to appreciate that many proteins associated with neurodegenerative diseases — including amyloid-β and tau in Alzheimer's disease and α-synuclein in Parkinson's disease — misfold catastrophically. Structural biologists call the entire family of misfolded

proteins (including PrP) amyloids. Amyloid-β clumps into whitish plaques, tau forms ribbons called tangles and α-synuclein creates fibrous deposits called inclusions.

A decade ago, these similarities prompted neuroscientist Mathias Jucker at the University of Tübingen in Germany to test whether injecting brain extracts containing misfolded amyloid-β into mice could seed an abnormal build-up of amyloid in the animals' brains. He found that it could, and that it also worked if he injected amyloids into the muscles[5]. "We saw no reason not to believe that if amyloid seeds entered the human brain, they would also cause amyloid pathology in the same way," says Jucker.

This didn't cause alarm at the time, because it wasn't clear how an amyloid seed from the brain of someone with Alzheimer's could be transferred into another person's body and find its way to their brain. To investigate that, what was needed was a group of people who had been injected with material from another person, and the opportunity to examine their brains in great detail, preferably when they were still relatively young and before they might have spontaneously developed early signs of Alzheimer's.

The CJD brains provided just that opportunity. Between 1958 and 1985, around 30,000 people worldwide received injections of growth hormone derived from the adrenal glands of cadavers to treat growth problems. Some of the preparations were contaminated with the prion that causes CJD. Like all prion diseases, CJD has a very long incubation period, but once it gets going it rages through the brain, destroying all tissue in its wake and typically killing people from their late 40s onwards. According to 2012 statistics[6], 226 people around the world have died from CJD as a result of prion-contaminated growth-hormone preparations.

Collinge had not set out to find a link with Alzheimer's — it emerged as part of routine work at the National Prion Clinic in London, which he heads, and where around 70% of all people in the United Kingdom who die from prion-related causes are now autopsied. The clinic routinely looks for signs of all amyloid proteins in these brains to distinguish prion disease from other conditions. It was thanks to this routine work that the cluster of unusual cases emerged of people who had clearly died of CJD, but who also had obvious signs of amyloid pathology in their grey matter and cerebral blood vessels.

As soon as he saw these brains, Collinge knew that he could get into stormy waters. Keen to strike a balance between warning of a possible public-health risk and causing unwarranted panic, he sketched a carefully worded press release that would go out from the National Prion Centre and set up hotlines for people who had been treated with growth hormone in the past. But no panic occurred: apart from one or two overwrought headlines, the news stories were fairly measured, he says. Only around ten people called the hotlines.

"ALL AMYLOIDS SHOULD BE CONSIDERED DANGEROUS UNTIL PROVEN SAFE."

For scientists, however, the paper was a red flag. "As soon as the paper came out we realized the health implications and started collecting slides and paraffin blocks from patients," says Jiri Safar, director of the National Prion Disease Pathology Surveillance Center at Case Western Reserve University in Cleveland, Ohio. Like other pathologists in countries where people had died of CJD associated with medical procedures, he rushed to check the centre's archives of autopsied brains to see if any of them contained the ominous amyloid deposits.

The answers are not yet in. Safar says that it has not proved easy to trace brain samples in the United States, but that he is working to do so with the National Institutes of Health and the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia. Charles Duyckaerts at the Pitié-Salpêtrière Hospital in Paris, France, has now examined brain tissues from around 24 patients and is likely to report the results later this year.

A further 228 cases of CJD were caused by transplantation of prion-contaminated dura mater — the membrane surrounding the brain and spinal cord — prepared from cadavers around the world. Dura-mater preparations were regularly used in brain surgery as repair patches until the late 1990s. For the study[2] published in January, Herbert Budka at the National Prion Diseases Reference Center at University Hospital Zurich and his colleagues examined the brains of seven such patients from Switzerland and Austria, and found that five had amyloid deposits in grey matter and blood vessels. In Japan, dementia researcher Masahito Yamada at Kanazawa University is making his way through a large number of such autopsy specimens and says that the 16 brains he has examined so far show signs of unusually high levels of amyloid deposition in cerebral blood vessels.

Yet such case studies can only ever provide circumstantial evidence that seeds of amyloid-β were transferred during the treatments. And they cannot entirely rule out the possibility that the treatments themselves — or the patients' original medical conditions — caused the amyloid pathology. More-conclusive evidence would come from checking whether the original growth hormone and dura-mater preparations contained infectious amyloid seeds, by injecting them into animals and seeing whether this triggers disease. Most of these preparations, however, have long since disappeared. Collinge has access to some original samples of growth hormone stored by the UK Department of Health, and he is planning to analyse them for the presence of amyloid seeds and then inject them into mice. That work will take a couple of years to complete, he says.

SEEDS OF DOUBT

There is another hitch: no one knows for sure what size and shape the amyloid seeds might be. Jucker is hunting for them in an unusual source of human brain tissue that has nothing to do with CJD. A team in Bonn has collected frozen samples from more than 700 people with epilepsy who were operated on over the past 25 years to remove tissue that was driving their seizures. "It is the best source of fresh human brain tissue available at the moment," says Jucker, who plans to scrutinize it carefully under the microscope for anything that might resemble tiny clumps or seeds of amyloid-β. The team also has records of the patients' cognitive skills, such as language and memory skills, before and at regular intervals after the operations. This should allow Jucker's team to correlate the presence of any amyloid-β seeds it finds with changes in the cognitive function of individual patients over time.

Scientists have shown that tau and α-synuclein can also seed pathological features in mice. In two studies[7,8] from 2012, scientists injected fibrils of α-synuclein into the brains of mice already engineered to develop some of the characteristics of Parkinson's disease. This triggered the early onset of some of the signs and symptoms of Parkinson's, and eventually killed the animals. A third study[9] showed that similar injections into normal mice caused some of the neurodegeneration typical of Parkinson's disease and the mice became less agile. In humans, α-synuclein would not necessarily turn out to be equally aggressive — mouse models of neurodegenerative diseases do not mimic human disease very closely — but scientists are taking the possibility seriously.

If the transmissibility hypothesis proves true, the implications could be severe. Amyloids stick like glue to metal surgical instruments, and normal sterilization does not remove them, so

amyloid seeds might possibly be transferred during surgery. The seeds might sit in the body for years or decades before spreading into plaques, and perhaps enabling the other pathological changes needed to induce Alzheimer's disease. Having amyloid plaques in cerebral blood vessels could be dangerous in another way, because they increase the risk that the vessel walls might break, leading to small strokes.

But if common medical procedures really increased the risk of neurodegenerative disorders, then wouldn't that already have been detected? Not necessarily, says epidemiologist Roy Anderson at Imperial College London. "The proper epidemiological studies have not been done yet," he says. They require very large and carefully curated databases of people with Alzheimer's disease, which include information about the development of symptoms and autopsy data. He and his team are now studying the handful of reliable databases that exist to tease out a signal that might associate medical procedures with Alzheimer's progression. The number of patients currently available may turn out to be too small to draw conclusions, he says, but a more definitive answer could emerge as the databases grow.

Faced with so much uncertainty, some researchers and public-health agencies have adopted a wait-and-see approach. "We are right at the beginning of this story," says Nicotera, "and if there is one message to come out right now it is that we need more work to see if this is a relevant mechanism." The CDC and the European Centre for Disease Prevention and Control in Solna, Sweden, say that they are keeping a cautious eye on the issue.

If further research does confirm that common neurodegenerative diseases are transmissible, what then? One immediate priority would be rigorous sterilization procedures for medical and surgical instruments that would destroy amyloids, in the way that extremely high temperatures and harsh chemicals destroy prions. Aguzzi says that funding agencies should put out calls now to researchers to develop cheap and simple sterilization methods. "It's not very sexy science, but it is urgently needed," he says. He also worries about the safety of researchers working with amyloids — particularly α-synuclein. "I have nightmares that someone in my lab may catch Parkinson's," he says. "While the story is in flux, our first duty is to protect lab workers."

STRAIN SEEKERS

The similarities between prions and other amyloids is throwing open other avenues of research. Prions can exist as distinct strains — proteins that have the same sequence of amino acids but misfold in different ways and have distinct biological behaviours[10], much as different strains of a pathogenic virus can be aggressive or weak. The outbreak of vCJD in the United Kingdom in the 1990s was traced to BSE-contaminated meat because the prion strain was the same in both.

Over the past few years, research in animals has shown that different strains of amyloid-β and α-synuclein exist[11,12]. And a landmark paper[13] in 2013 reported that strains of amyloid-β with different 3D structures were associated with different disease progression in two people with Alzheimer's. Structural biologist Robert Tycko, who led the work at the National Institute of Diabetes and Digestive and Kidney Diseases in Bethesda, Maryland, is now looking at many more brain samples from such patients.

Knowing the structures of pathological forms of amyloid seeds should help to design small molecules that bind to them and stop them doing damage, says biophysicist Ronald Melki at the Paris-Saclay Institute of Neuroscience, who works on α-synuclein strains. His lab is designing small peptides that target the seeds and mimic regions of 'chaperone' molecules, which usually bind to proteins and help them to fold correctly. Melki's small peptides mimic these binding regions, sticking to the amyloid proteins to stop them from aggregating further.

In the research community, much of the agitation in response to Collinge's paper boils down to semantics. Some scientists do not like to use the word 'prion' in connection with the amyloids associated with common neurodegenerative diseases, or to describe any of their properties as 'prion-like' — because of its connotation of infectious, deadly disease. "The public has this perception of the word 'prion'," says Alzheimer's researcher Brad Hyman at Harvard Medical School in Boston, Massachusetts, and this matters, even if their ideas are wrong. "One of my patients told me that she wasn't getting any hugs any more from her husband who had read about the case in the media — that made me sad," he says.

Others, however, feel that it is helpful to consider prions and other amyloids as being part of a single spectrum of conditions involving proteins that misfold and misbehave. It means that researchers studying prion diseases and neurodegenerative diseases, who until recently had considered their disciplines to be separate, now find themselves tackling shared questions.

Both fields are wary of raising premature alarm, even though they wonder what the future will bring. Jucker, only half-jokingly, says he could imagine a future in which people would go into hospital every ten years or so and get the amyloid seeds cleared out of their brains with antibodies. "You'd be good then to go for another decade." ■

"THERE IS NO CONCLUSIVE EVIDENCE THAT SEEDS OF AMYLOIDS CAN TRANSMIT ACTUAL DISEASE."

AFP/STRINGER

Alzheimer's, primarily a disease of elderly people, has complex roots.

Alison Abbott *is* Nature*'s senior European correspondent.*

1. Jaunmuktane, Z. *et al. Nature* **525,** 247–250 (2015).
2. Frontzek, K., Lutz, M. I., Aguzzi, A., Kovacs, G. G. & Budka, H. *Swiss Med. Wkly* **146,** w14287 (2016).
3. Prusiner, S. B. *Science* **216,** 136–144 (1982).
4. Chiesa, R. *PLoS Pathog.* **11,** e1004745 (2015).
5. Meyer-Luehmann, M. *et al. Science* **313,** 1781–1784 (2006).
6. Brown, P. *et al. Emerg. Infect. Dis.* **18,** 901–907 (2012).
7. Mougenot, A. L. *et al. Neurobiol. Aging* **33,** 2225–2228 (2012).
8. Luk, K. C. *et al. J. Exp. Med.* **209,** 975–986 (2012).
9. Luk, K. C. *et al. Science* **338,** 949–953 (2012).
10. Aguzzi, A., Heikenwalder, M. & Polymenidou, M. *Nature Rev. Mol. Cell Biol.* **8,** 552–561 (2007).
11. Cohen, M. L. *et al. Brain* **138,** 1009–1022 (2015).
12. Peelaerts, W. *et al. Nature* **522,** 340–344 (2015).
13. Lu, J-X. *et al. Cell* **154,** 1257–1268 (2013).

COMMENT

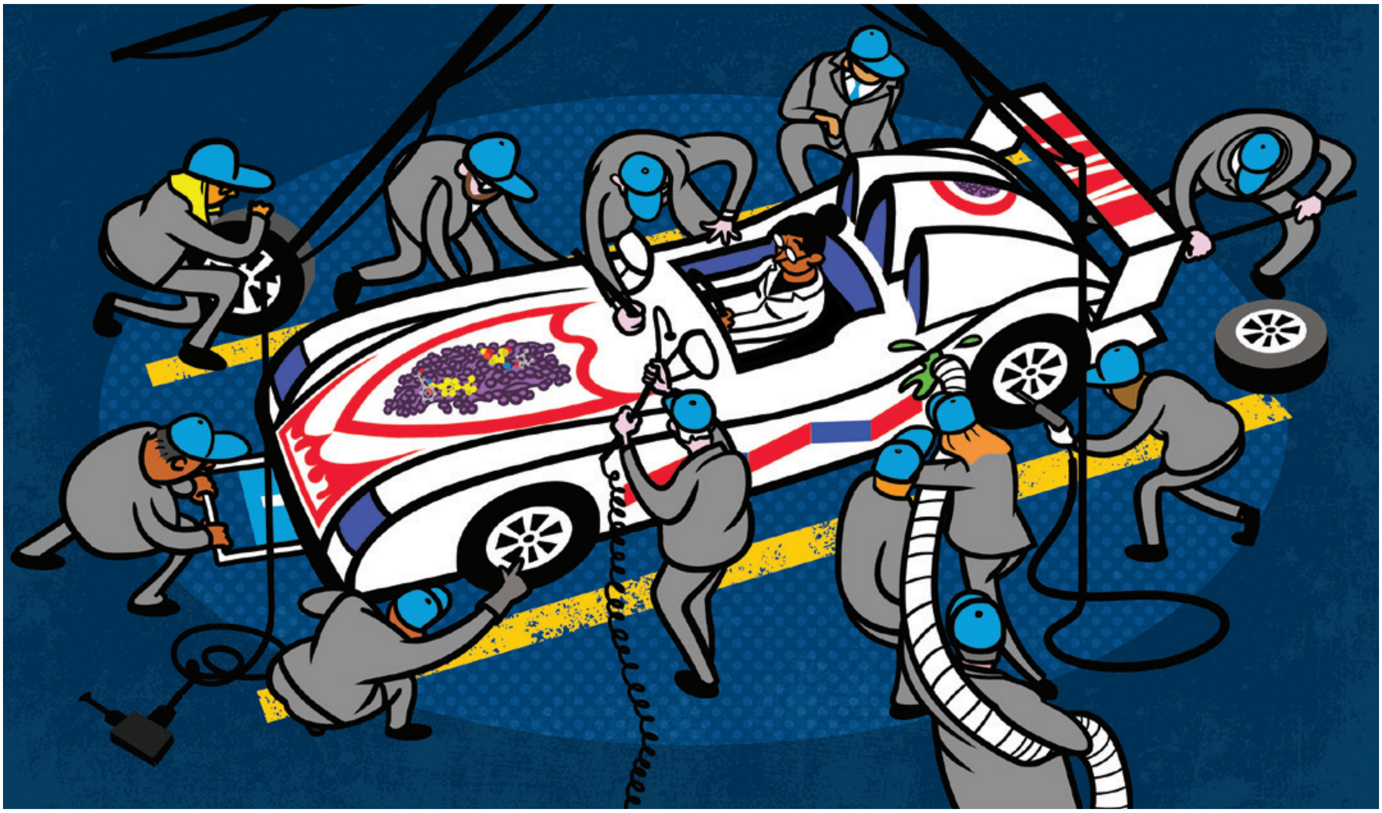
MUSEUMS To whom do collected objects belong, why and what for? **p.302**

ENTOMOLOGY A hymn to Vladimir Nabokov's powers of observation **p.304**

SUSTAINABILITY Cuts to environment and climate jobs are short-sighted **p.305**

FOOD IPBES must account for contribution of biodiversity to agriculture **p.305**

ILLUSTRATIONS BY PETE ELLIS



Team up with industry

Combining commercial and academic incentives and resources can improve science, argues **Aled Edwards**.

The scientific community is bustling with projects to make published results more reliable. Efforts are under way to establish checklists, to revamp training in experimental design, and even to fund disinterested scientists to replicate others' experiments. A more efficient strategy would be to rework current incentives to put less emphasis on high-impact publications, but those systems are entrenched, and public funders and universities are ill-prepared for that scale of change.

To catalyse change, industry must step up to the plate. I have learned this first hand, as head of the Structural Genomics Consortium (SGC), a research charity funded by business, government and other charities. If more companies contributed funds and expertise to efforts such as ours, I believe it would create a system that rewards science that is both cutting-edge and reproducible.

Here I share my experience of running a collaboration between academic and industrial researchers over the past 12 years, and distil the principles that have made for reliable research.

THEN AND NOW

In the late 1990s, several drug companies concluded that academic researchers were only half-heartedly pursuing structural information about human proteins, information crucial to the design of new drugs. They proposed pooling resources with government bodies to fund this work. And so, in 2004, drug company GlaxoSmithKline, the major UK biomedical funder the Wellcome Trust and science funders from Canada launched the SGC. Our task? To generate human protein structures and place them in the public domain.

The motivation varied. Public and charitable funders emphasized scientific novelty and high-quality publications. Industry wanted research relevant to drug-discovery efforts. For example, companies insisted on structures of human proteins rather than those from other species, ▶

even though many of these would have been of great scientific interest. All funders demanded quantifiable milestones, unrestricted use of data and reagents, and the right to withdraw support from the project if it underperformed (see 'Milestones exceeded').

The number of drug companies participating has grown to eight. The SGC currently disburses more than US$20 million to 250 scientists in 6 dedicated laboratories. Our researchers include principal investigators, postdocs, technicians and graduate students. At any given time, as many as 50 scientists in industry collaborate with SGC scientists. Projects now extend beyond protein structures to the discovery of chemical probes (small molecules useful for studying protein function), and to hospital-based collaborations that explore the effects of these tools in cells from patients. All the resources we create are non-proprietary and readily available.

"Done well, industry–academic collaborations could recreate the engine that powered successful enterprises."

In 2010, as the SGC began to attract the interest of pharmaceutical companies that had never before participated in consortia dedicated to producing public information, I asked, "Why us?", hoping to hear about our amazing scientific intellect. The answer was more prosaic: "We can repeat your work." Back then, before reports emerged that fewer than half of biomedical papers could be reproduced (L. P. Freedman *et al. PLoS Biol.* **13,** e1002165; 2015), that did not make much sense to me. It does now.

Ironically, it was our work with industry that had helped make our research so reliable in the first place. Our industrial colleagues helped to design processes to increase the chances that they could depend on our work. From the outset, we were operating within a system in which continued funding was tied to research that proved useful.

EIGHT PRINCIPLES

In my opinion — shared by my long-standing industrial collaborators — several mutually reinforcing factors characterize this research system. Each is essential. It is the combination, rather than any single principle, that is key.

Require full commitment, and reward efficiency. Focus and organization are necessary for success. As a precondition for receiving funding, scientists agree to dedicate all their research time to the project. This is the aspect in which the SGC differs most from other schemes involving academic partnerships with industry. For participating scientists, the advantage is efficient, predictable funding: as long as scientists achieve their milestones, expenses are covered without the need for grant applications. It also incentivizes scientists to innovate. If they achieve their targets without using all their available funds, they can use the remainder to pursue their curiosity.

Define objectives that cannot be achieved with current technology. Many scientists and public funders believe that formal milestones — a requirement for industry investment — are the antithesis of discovery research. The solution is to create 'stretch goals'. In our case, funders provided a list of 2,000 human proteins and a directive to solve the structures of 350 — knowing full well that that goal was not achievable with the technologies of the day. This worked. To meet these and subsequent milestones, SGC scientists developed new methods and have published more than 800 peer-reviewed papers (of which 60 have appeared in *Nature* and its eponymous sister journals).

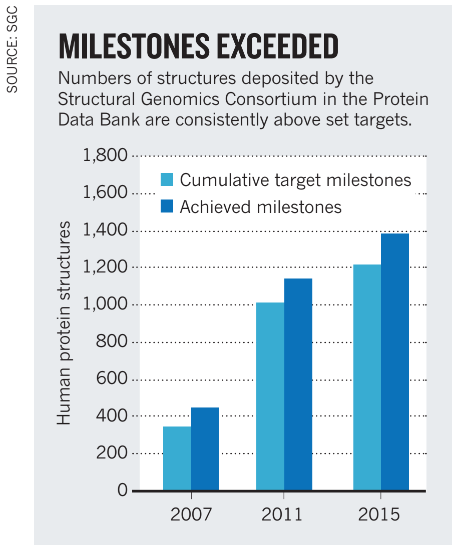
Establish clear quality criteria and make them public. Milestones must be unambiguous, or else they can be gamed. In our case, we asked a group of independent experts from academia and industry to craft quantitative criteria to judge research outputs. For instance, we defined how different a protein sequence needed to be from others in the Protein Data Bank to count as 'novel', and specified acceptable levels of precision and error. All these criteria were published on our website (see go.nature.com/4qncnj). SGC scientists knew that protein structures would not be 'counted' unless they met or exceeded the publicly available criteria.

Mandate data sharing. Mechanisms for sharing progress and enforcing transparency promote reproducibility. We rolled out electronic lab notebooks in 2004. There was some initial resistance, but scientists quickly realized that ready access to information from their colleagues helped their own research. Electronic lab notebooks make it easier to document and disseminate detailed experimental procedures, and also reduce the risk that researchers will cherry-pick data.

Subject work to independent oversight before public release. Participants should not be the judges of whether they have met a milestone. We created an independent advisory board of academic and industry scientists to assess the quality of each structure, probe, or other tool before it was released into the public domain. This external body prevents us from loosening the quality criteria for research outputs if achieving the original goals turns out to be harder than expected.

This oversight was more tedious and humbling than we were used to. Consortium scientists had to prepare documents defending how each criterion had been met, and external scientists who intend to use the potential tool tend to be tougher judges of robustness than the average peer reviewer.

Over the years, however, external vetting became simply part of our practice. Without a doubt, it contributed to our reputation for reproducible and meaningful research. On two occasions, SGC scientists were convinced that a chemical probe (see go.nature.com/sediul) had met our quality criteria, only to have an oversight committee raise what turned out to be legitimate concerns, which were subsequently addressed.



Enshrine public ownership for all research outputs. The value of foundational science is greater when more people use the research, so materials must be disseminated as freely as possible. We avoided legal encumbrances such as formal Material Transfer Agreements (MTAs) where possible, developed 'click-wrap' MTAs where necessary, and encouraged vendors to provide all our reagents without restriction. The result is that many scientists outside the SGC use our materials. Currently, the SGC contributes 5% of all the plasmids deposited at the repository Addgene; last year, vendors received more than 5,000 orders for chemical probes generated by the SGC.

Ensure that industry and academic scientists collaborate. Industry must provide expertise as well as funds. Collaboration with industry scientists engenders a shared desire to succeed, and creates a sense of ownership of a project. The different motivations also create productive tension. For example, scientists in academia have strong incentives to publish rapidly. Unfortunately, this can lead to the publication of stories that are true only under narrowly defined conditions. By contrast, industry scientists push for validation using a range of orthogonal experiments; these alternative ways of evaluating the same research tool ensure that the results are broadly useful. Publications and release must happen by consensus. By balancing these desires, we achieve an optimal combination of innovation, timely dissemination and reproducibility.

Create an active governing body. Many academic projects, including a number that are co-funded by industry, create 'friends and family' governance bodies that advise, but do not govern. The SGC opted for a different model. Our board includes senior executives from every major funder. It can halt a project, change leadership and direct strategy. Any meaningful change to the project budget requires unanimous board approval. Because of this responsibility, in-person attendance has been nearly 100% for every quarterly board meeting for more than 12 years. A productive public–private-sector tension also plays out at the board. If industry requests a delay in releasing research results to further review data quality, public funders balance the scientific merits of this request with the need to disseminate the output rapidly.

EXPAND THE MODEL

Adherence to these principles builds an ecosystem that supports reproducible, innovative research. Scientific publications are no longer the sole units of achievement; reaching predefined milestones and making useful tools are also key to continued funding.

This approach cannot work for every scientific question or for every scientist. But there are many important scientific questions in which the long-term needs of industry and the talents of academic researchers are aligned. Here the research effort could be readily focused within a single, organized project where unrestricted access to the research results would benefit everyone. The oil and gas sector might pool resources to develop novel remediation technologies; the forestry sector might create a consortium to explore cellulose-degrading technologies; the renewables sector might collaborate to identify better energy-storage technologies.

Such public–private collaborations are in industry's interest. Companies need foundational science to innovate and for long-term profitability. In the past, researchers at BASF, Dupont and Bell Labs invented technologies that transformed science, business and daily life — from transistors to radar to synthetic fertilizers — and they won Nobel prizes. Today, support for internal company research has been slashed, and academic research is not filling the gap because it is unpredictable, unfocused and, often, unreliable. Done well, industry–academic collaborations could recreate the engine that powered successful enterprises.

Such partnerships should be more widespread. They can advance important areas of research. Equally important, the research community gains knowledge and tools, and the practices that make for robust science might diffuse beyond the collaborations to raise the quality of science overall. ■

Aled Edwards *is chief executive of the Structural Genomics Consortium.*
e-mail: aled.edwards@utoronto.ca

MÜTTER MUS. OF THE COLLEGE OF PHYSICIANS OF PHILADELPHIA

Bones from the US Civil War, on display at the Mütter Museum in Philadelphia, Pennsylvania.

MUSEUMS

Ethics of exhibition

David Hurst Thomas explores the controversies over collections of human remains and plundered artefacts.

Three months before the 1941 attack on Pearl Harbor, anthropologist Margaret Mead published a prescient piece in *Natural History* entitled 'Museums in the Emergency'. Mead found the US citizenry "suspicious of every means of communication" — except museums. She attributed this remarkable faith to the museum practice of asking, "Is this true?" rather than, "Will this make a hit?" A stubborn insistence on truth, Mead believed, could keep the museum as "a place in which [people] can renew their trust in science and in democracy".

Two books look at truth and trust in the museum world. In *Bone Rooms*, historian Samuel Redman tracks the evolving role of collections of human remains and their public display in framing issues of race. Cultural sociologist Tiffany Jenkins crafts a spirited read in *Keeping their Marbles*, with much to offer regarding the genesis of the world's great museum collections. She transports the reader from the Napoleonic campaigns that stocked the Louvre in Paris with Egyptian treasures to British imperialists funnelling global booty to London's British Museum. Many countries now want their treasures back; Nigeria, for instance, wants the return of its bronzes, taken when the British Army flattened the then-kingdom of Benin in the late nineteenth century. Both books explore the question of who owns the past, with remarkably different answers. *Keeping their Marbles* advocates maintaining the finders-keepers mentality that created the museum collections. *Bone Rooms* argues that human remains were sometimes inappropriately acquired in the name of science, and that meaningful steps must be taken to redress the balance.

Redman documents the US 'skull wars' of the late nineteenth century, when museums competed to collect human skulls, whole skeletons, mummies and fossils. Battlefield casualties became fair game, as

Bone Rooms: From Scientific Racism to Human Prehistory in Museums
SAMUEL J. REDMAN
Harvard University Press: 2016.

Keeping Their Marbles: How the Treasures of the Past Ended Up in Museums and Why They Should Stay There
TIFFANY JENKINS
Oxford University Press: 2016.

did archaeological sites, Native American cemeteries and indigenous people unfortunate enough to pass away at a World's Fair. Even isolated body parts were accessioned, as bizarrely illustrated by the Civil War veteran who found his own amputated arm in the Army Medical Museum in Washington DC.

Bone Rooms details the nascent views of racial science that evolved in US natural history, anthropological and medical museums. These debates spilled into public museum spaces, arraying human bodies in sometimes controversial, even macabre, exhibits. Redman effectively portrays the remarkable personalities behind them, particularly pitting the prickly Aleš Hrdlička at the Smithsonian Institution in Washington DC against ally-turned-rival Franz Boas at the American Museum of Natural History in New York City. Debates over racial science should have evaporated when Boas conclusively demonstrated that language, culture and biology ('race') are independent — the premise of modern anthropology. But the myth of scientific racism (typically involving questionable concepts such as 'racial essence' and 'racial genius') persisted for a century more, due in part to Hrdlička's powerful influence over widely attended exhibitions that promoted the now-discredited idea that races are immutable, with evolutionary change transpiring only within (not between) human races.

Bone Rooms also highlights ethical concerns over collecting, curating and exhibiting human remains that simmered in the 1930s and boiled over after the Vietnam War. Native Americans increasingly expressed shock at the tens of thousands of ancestral skeletons appropriated without their permission or, usually, knowledge. The Native American Graves Protection and Repatriation Act of 1990 required museums to inventory holdings of human remains (plus potential sacred, patrimonial and funerary artefacts), then to consult tribes about cultural affiliations and time and manner of repatriation. Although most human remains stayed in storage, tens of thousands were repatriated. A few hotly disputed cases linger on, including Kennewick Man (an 8,400-year-old skeleton now held in limbo at the Burke Museum of Natural History and Culture in Seattle, Washington). Redman concludes (correctly, in my view) that, on balance, repatriation programmes are "successful steps forward".

Keeping their Marbles views museum truth and trust differently. Jenkins does an excellent

job of portraying the extreme reactions elicited by repatriation conversations — from the smug 'we-stole-it-fair-and-square' to the angst parodied by historian Elazar Barkan as 'performance guilt' (in which "leaders theatrically say sorry for acts from the past for which they had no responsibility"). Although granting that great museum collections "were wrenched from their original contexts by means that often amounted to theft", Jenkins bristles at returning items. Rather, she stresses three principles — preservation, truth and access — that determine what is best for objects, scholars and the public. "The mission of museums," she argues, "should be to acquire, conserve, research, and display their collections ... That is all and that is enough."

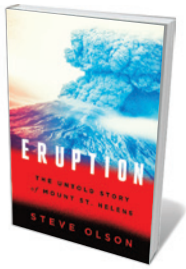
I don't believe that is enough — particularly with respect to human remains. As a practising dirt archaeologist, I still on occasion excavate human burials. As a museum curator, I sometimes exhibit human remains. But as a museum-based researcher, I acknowledge my responsibilities to consult meaningfully with relevant descendant and stakeholding communities — and listen to what they say. Jenkins is correct that repatriation will render some human remains unavailable for research. The public display of certain human remains is likewise often inappropriate. These are limitations that many of us accept today.

The alternative — the free-ranging, science-*über-alles* mentality articulated in *Keeping Their Marbles* — reprises the cavalier attitudes towards communities of descendants that characterized Americanist archaeology for most of the nineteenth and twentieth centuries. That sordid legacy, which necessitated reburial and repatriation legislation in the first place, seems particularly inappropriate for the responsible practice of twenty-first-century science. What of Margaret Mead's belief in modern truth and trust? Today's headlines target different 'Museums in the Emergency', from the systematic looting of the National Museum in Baghdad to the Islamist terrorist group ISIS taking sledgehammers to Syrian antiquities. The prominent Syrian scholar Khaled al-Asaad was beheaded by ISIS for refusing to disclose where ancient treasures from Palmyra had been hidden for safe keeping. The right of museums to hold and display collections is today contested at every turn.

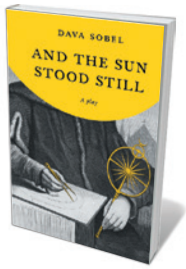
Modern museums have multiple meanings, objectives and constituencies. But one thing is certain: nowhere is there now a museum where all people "can renew their trust in science and in democracy". No matter where the Parthenon Marbles end up, that museum world has become an artefact of the past. ■

David Hurst Thomas *is a curator of anthropology at the American Museum of Natural History, New York City.*
e-mail: thomasd@amnh.org

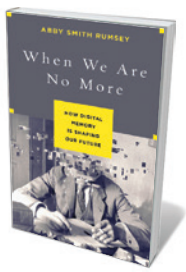
Books in brief



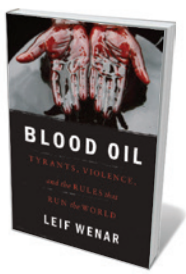
Eruption: The Untold Story of Mount St. Helens
Steve Olson W. W. NORTON *(2016)*
When Mount St Helens in Washington state erupted in 1980, it woke the nation from political and economic torpor. The huge sideways blast — a pyroclastic flow — killed 57, triggered the largest landslide ever recorded and spewed ash over 11 states and several Canadian provinces. Steve Olson intercuts stories of victims including David Johnston, the volcanologist who was monitoring the explosion, with an account of its impact on science — such as popularizing the use of lidar. With 1,500 potentially active volcanoes worldwide, this is an urgent reminder of the need for advances in the field.



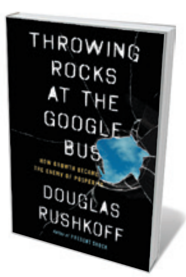
And the Sun Stood Still: A Play
Dava Sobel BLOOMSBURY *(2016)*
The centrepiece of science writer Dava Sobel's acclaimed 2011 history *A More Perfect Heaven* (Bloomsbury) is a dramatized telling of a crucial meeting: the 1539 encounter between Nicolaus Copernicus and German mathematician Georg Joachim Rheticus, who would broker the publication of the Polish astronomer's great treatise on heliocentrism, *De Revolutionibus*. Now reworked as a play, Sobel's imaginative exploration of how Rheticus convinced the "starry canon" to air his theory is a revelation of world-shifting science illuminating the human mind, leavened with a sparkling immediacy.



When We Are No More: How Digital Memory Is Shaping Our Future
Abby Smith Rumsey BLOOMSBURY *(2016)*
A door is opening on a frightening prospect: the future of history. So notes scholar Abby Smith Rumsey in this erudite treatise on how the digitization of archival technology makes it all too easy to rewrite our cultural past. She analyses our journey in recorded memory, interweaving neuroscience with a history of the archive, and ranging from classical mnemonic devices to the collective amnesia that can follow the destruction of libraries. Books, she shows, are "memory machines" that we have learned to manage. Digitized data *in toto* is a different beast — and one bucking under our attempts at control.



Blood Oil: Tyrants, Violence, and the Rules That Run the World
Leif Wenar OXFORD UNIVERSITY PRESS *(2016)*
Petroleum is truly globalized. But for exporting countries such as Nigeria, high-demand raw materials can be a "resource curse", linked to political corruption and socio-economic inequality (see J. Vidal *Nature* **482,** 306; 2012). In this straight-talking manifesto, philosopher Leif Wenar draws on economics and political science to call for a rethink on global supply chains. Clean trade policies to protect public property and accountability are needed, he argues, if poorer nations are to achieve resource sovereignty and Western importers are to stop buying blindly into oppressive regimes.



Throwing Rocks at the Google Bus
Douglas Rushkoff PORTFOLIO *(2016)*
Technology writer Douglas Rushkoff delivers an incisive analysis of digitized culture in this shrewd study of the economic rot at its heart. Issues such as the corporate growth model and "platform" monopolies are, he shows, threatening the public good. He suggests that the rage of protestors who attacked shuttle buses carrying Google employees in 2013 would be better channelled into "digital distributism" — an economy that hinges on democratic ownership of the means of production, cooperatives and genuine sharing. **Barbara Kiser**

ENTOMOLOGY

Nabokov’s scientific artistry

Vladimir Lukhtanov delights in a treatise on the luminary’s contribution to biology.

One of Vladimir Nabokov’s drawings of the undersurfaces of butterfly wings.

Vladimir Nabokov’s influence on Russian and English literature and language is assured. Many people also know of the novelist’s lifelong passion for butterflies. But his notable contributions to the science of lepidopterology and to general biology are only beginning to be widely known.

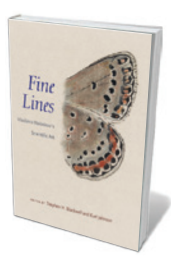
Nabokov was no amateur entomologist. He served for six years as curator of the butterfly collection at Harvard University’s Museum of Comparative Zoology in Cambridge, Massachusetts, and published a dozen papers on taxonomy — the description and classification of organisms — that remain important. His observations on butterfly morphology have stimulated breakthrough research in evolutionary biology. Several of his original biogeographic hypotheses have been confirmed in the past few years. *Fine Lines*, a collection edited by Stephen Blackwell and Kurt Johnson, explains the importance of Nabokov’s scientific work and traces its influence on his novels.

The book begins with 154 of Nabokov’s black-and-white and colour drawings of butterflies’ fine anatomical structures. Most represent the European, Asian and American species of the ‘blues’ of the tribe Polyommatini, Nabokov’s favourite group. Ten essays follow, by prominent researchers including evolutionary biologist James Mallet, current Harvard butterfly curator Naomi Pierce and lepidopterist Robert Pyle, explaining the interplay of science and art in Nabokov’s writings. *Fine Lines* clearly demonstrates the significant impact that science had on Nabokov’s evolution as a writer.

The decision to open the book with the drawings is a masterstroke. They illustrate one of the most important aspects of Nabokov’s creativity — his tremendous attention to details, described with scrupulous precision. In his novels, he seamlessly marshals minutiae — impressions, passing fancies, ideas — to create a universe strongly rooted in observation. The particular or apparently trivial was, for him, always worth probing. In his entomological studies, he analysed fine, nearly invisible, dots on the wings of New and Old World butterflies to hint at what may have happened on Earth millions of years ago. With no palaeontological data, Nabokov speculated that North and then South America were populated by five waves of butterflies migrating from Asia (V. Nabokov *Psyche* **52,** 1–61; 1945) — a picture confirmed by DNA analysis almost 70 years on (R. Vila *et al. Proc. R. Soc. B* **278,** 2737–2744; 2011).

This pointillism is harder than it seems. Piling up millions of elements can easily end in chaos; to create a picture, one needs to understand the nature of these elements and to be able to choose between them. The core of scientific drawing differs greatly from photography in focusing on the heart of the matter and avoiding unnecessary details. This is important for science, and no less for art. Both have the same central goal — to reveal an unknown or invisible essence of things. That is one of the main points of *Fine Lines*.

Yet science and art diverge in their communication. In science, the ability to convey the idea properly and simply is a matter of special talent, but almost everyone can learn to do it. Not so in art. Nabokov’s drawings are scientifically perfect, but also staggeringly fine aesthetically. They show how the merging of content and form in art conveys ideas wonderfully. However, even the most wonderful idea becomes banal if artistry is lacking.



Fine lines: Vladimir Nabokov’s Scientific Art
EDITED BY STEPHEN H. BLACKWELL AND KURT JOHNSON
Yale University Press: 2016.

The personal, artistic and scientific aspects of Nabokov’s life were tightly intertwined. As one of the book’s essayists, science writer Dorion Sagan, concludes, nature and art were a continuum for him: “the distinct but equally necessary paths of art and science seem to scale opposite sides of the same majestic mountainscape”.

Nabokov’s fiction is permeated by science, as *Fine Lines* amply reveals. He was a master in the use of motif and symbols. In his novel *Lolita* (Olympia, 1955), for instance, the town Lepingville is named after ‘lepping’, butterfly hunters’ slang for chasing butterflies, and Elphinstone after *Elphinstonia*, a subgenus in the white butterfly genus *Euchloe*. The fictional play-within-the-novel, *The Enchanted Hunters*, is built almost entirely on symbols associated with butterflies. Diana, its protagonist, is both the virgin goddess of hunting and a butterfly species (*Speyeria diana*). In his essay, Nabokov scholar Brian Boyd reveals that Edusa Gold, who directs the play, is an echo of *Colias edusa*, an old but preocccupied name for the clouded yellow (now *Colias croceus*). I can add that her sister Electra Gold was named after *Colias electra*, an unavailable name for the African clouded yellow, now *Colias electo*. That these names are effectively hidden — no longer in use, but buried in lists of unavailable scientific epithets — chimes with the secrecy in this controversial novel.

I prepared this review at the Nabokov House Museum in St Petersburg, Russia. While there, I discovered in the Nabokov family’s copy of *An Illustrated Natural History of British Butterflies and Moths* by Edward Newman (William Glaisher, 1870) that Nabokov had, as a child, coloured in the black-and-white image of the clouded yellow with remarkable accuracy. As zoologist Victor Fet describes in *Fine Lines*, Nabokov’s childhood concentration on butterfly collecting and drawing effectively provided very specific training in memory and paying attention, as well as that focus on minute detail.

Few have so beautifully and meaningfully meshed serious scientific endeavour with artistic brilliance, visual and verbal. *Fine Lines* helps us to understand the phenomenon of creativity, without which neither good science nor true art can exist. ■

Vladimir Lukhtanov *is leading research scientist at the Zoological Institute of the Russian Academy of Sciences in St Petersburg, and professor of entomology at Saint Petersburg State University.*
e-mail: lukhtanov@mail.ru

Correspondence

Short-sighted to cut environment posts

Australia's premier government-funded science agency, CSIRO, is shedding jobs in environmental science — 100 scientists from the climate-sciences division alone face imminent job loss. The perception that environmental research is unprofitable has already rendered it victim to four years of government cuts. Evidently, the promise of significant savings from science-based environmental remediation has yet to resonate with decision-makers.

Other major CSIRO job losses affect research on sustainable management of the nation's terrestrial and aquatic ecosystems and biodiversity. James Cook University in Queensland intends to axe 25% of its academic staff in the environmental sciences. Effective research organizations such as Land and Water Australia have been disbanded. And past cuts have already translated into weakened environmental regulation and management.

Australia is vulnerable to climate-change effects and has rapid population growth. It leads the world in recent extinctions of terrestrial mammals, and has vast areas that are in urgent need of restoration after widespread intensive land clearing, livestock overgrazing and mining.

Anti-environment policies and further destruction of Australia's research capability are threatening to destroy its priceless natural heritage.

David Lindenmayer *The Australian National University, Canberra, Australia.*
david.lindenmayer@anu.edu.au

China draws lines to green future

To conserve the strategic integrity of its environment, China has drawn up a system of 'Red Lines'. These denote the total minimum areas of various land-use types nationally and regionally, without specifying their exact locations.

Coming after Red Lines that were created to protect cropland and forest habitats, the latest Red Line will safeguard China's vast biodiversity, environmental resources and ecosystem services. This could consolidate the shift in the country's environmental strategy, which is moving away from networks of protected areas and short-term ecological restoration towards longer-term conservation of entire landscapes.

We propose that the area marked by the latest Red Line (see go.nature.com/na6ry6; in Chinese) should equal at least 496 million hectares. This would incorporate the areas covered by China's existing nature-reserve network and three recent, overlapping, landscape-scale conservation schemes. Within these, 'priority biodiversity conservation areas', 'important ecosystem-function areas' and 'key ecosystem-service function areas' have been designated for flood protection, erosion control, biodiversity conservation and ecosystem-service provision.

Direct government payments for ecosystem-service provisions and adjustments to imbalances in designations, particularly in China's eastern provinces, should underpin this initiative.

Sang Weiguo *Minzu University of China; and Institute of Botany, Chinese Academy of Sciences, Beijing, China.*
Jan C. Axmacher *University College London, UK.*
j.axmacher@ucl.ac.uk

Six principles for EU peer review

The European Union Agencies Network for Scientific Advice (EU-ANSA) is a group of 11 EU agencies that provides scientific information for institutions and national authorities in Europe. It has recently assessed its peer-review practices and drawn up guidelines that we hope will support the agencies' work and contribute to the debate on changes to peer-review methods, particularly when there are implications for policymaking.

The context in which EU agencies provide scientific advice and technical support is highly specific. It requires a special approach to peer review that addresses particular challenges, such as ensuring reviewers' motivation, independence and international perspective.

The EU-ANSA collaboration has outlined six guiding principles (see go.nature.com/bqb7jp). These focus on: defining the process so that it is widely accepted; recruiting high-quality expertise; achieving credibility and avoiding bias using a transparent process; the provision of adequate resources; the availability of adequate technical support; and integration with established agency processes for ensuring good quality and performance in work.

William Cockburn *European Agency for Safety and Health at Work, Bilbao, Spain.*
Hubert Deluyker *European Food Safety Authority, Parma, Italy.*
cockburn@osha.europa.eu

Hasty publication compromises rigour

The period between submitting a paper and its publication can sometimes exceed the time invested in achieving the results (see *Nature* **530**, 148–151; 2016). Meanwhile, acquiring funding still hinges on clocking up publications in journals with high impact factors. These pressures pose a difficult choice for researchers, especially those in the early stages of their careers.

To compete for a faculty position or an independent grant, a PhD student must publish a handful of papers and then keep the ball rolling as a postdoc with at least one paper every year. The importance of publishing fast is hammered home during this period when, in our view, the emphasis should instead be on the reproducibility of results. In this battle for survival and growth, it is scientific rigour that might pay the price.

Even improvements in performance indicators — using citation counts, for example — are of little help to an early-career researcher. We need to move away from the 'publish or perish' ethos and towards incentives that reward scientific quality.

Shraddha Madhav Karve *Indian Institute of Science Education and Research, Pune, India.*
Madhur Mangalam *University of Georgia, Athens, USA.*
madhur.mangalam@uga.edu

Biodiversity central to food security

At the fourth plenary session of the Intergovernmental Panel on Biodiversity and Ecosystem Services (IPBES) last month, the thematic assessment on sustainable biodiversity use was referred for a second scoping by experts. I suggest that the new analysis needs to include biodiversity's contribution to ecosystem services that are essential to agricultural sustainability and food security.

Agricultural ecosystems are directly linked to human and environmental health (D. Tilman and M. Clark *Nature* **515,** 518–522; 2014). They are essential to the IPBES' ambition to influence progress on the United Nations Sustainable Development Goals. Furthermore, by 2050, the food supply for 9.6 billion people will depend on the sustainable use of agricultural biodiversity and its multiple ecosystem services (see go.nature.com/7ympnb).

A narrow scope that focuses on harvesting wild, uncultivated species will fail to capture biodiversity's importance to ecosystem services. Instead, we need a systems-based approach (see J. Liu *et al. Science* http://doi.org/627; 2015).

Fabrice DeClerck *Bioversity International — CGIAR, Montpellier, France.*
f.declerck@cgiar.org

Dry-season greening of Amazon forests

ARISING FROM D. C. Morton *et al.* *Nature* **506,** 221–224 (2014); doi:10.1038/nature13006

Evidence from ecological studies[1,2], eddy flux towers[3–5], and satellites[3,6] shows that many tropical forests 'green up' during higher sunlight annual dry seasons, suggesting they are more limited by light than water. Morton *et al.*[7] reported that satellite-observed dry-season green up in Amazon forests is an artefact of seasonal variations in sun-sensor geometry. However, here we argue that even after artefact correction, data from Morton *et al.* show statistically significant increases in canopy greenness during the dry season. Integrating corrected satellite with ground observations indicates that dry-season forest greening is prevalent in Amazonia, probably reflecting large-scale seasonal upregulation of photosynthesis by canopy leaf dynamics. There is a reply to this Brief Communication Arising by Morton, D. C. *et al. Nature* **531,** http://dx.doi.org/10.1038/nature16458 (2016).

Variations in sun-sensor geometry induce artefacts in remotely sensed vegetated surfaces[8]. Satellite studies thus typically use models to correct artefacts (for example, Moderate Resolution Imaging Spectroradiometer (MODIS) leaf area index[9], and multiangle implementation of atmospheric correction (MAIAC) enhanced vegetation index[10] (EVI)) or compositing algorithms designed to minimize artefacts (standard MODIS EVI[11]). Morton *et al.*[7] used a modelling approach to correct MODIS satellite data, which they state removed seasonal changes in surface reflectance, and redefined debates over how climate controls forest productivity in the Amazon. Setting aside arguments that the remote sensing analysis by Morton *et al.* is faulty[12], we take their correction[7] at face value, and ask two questions.

First, we ask whether the corrected results support their core conclusion that dry-season green up, previously observed by MODIS EVI, is eliminated. The hypothesis that Amazon forests green up in the dry season[3] can be rigorously evaluated by formal statistical tests. Morton *et al.*[7] showed that their correction reduces estimated dry season green up, ΔEVI (the EVI change during the dry season, ΔEVI = October EVI − June EVI; figure 3 in ref. 7 and Fig. 1). As the corrected mean ΔEVI was smaller than an *a priori* estimate of error for individual EVI observations, they concluded that the corrected mean ΔEVI was indistinguishable from zero. We find that this comparison, however, is not appropriate for assessing whether corrected EVI can resolve a basin-wide green up. The correct comparison, of mean ΔEVI to the error of the mean of the whole population of observations, is accomplished with standard statistical tests that lever the probability theory 'law of large numbers'[13]. For example, the 95% confidence interval[13] for basin-wide mean of corrected ΔEVI significantly excludes zero (Fig. 1). Alternatively, the corrected ΔEVI distribution[7] can be compared to the binomial distribution generated by the null hypothesis that pixels are equally likely to exhibit positive or negative ΔEVI (Fig. 1), which is analogous to treating 'green up' or 'brown down' as the outcome of the flip of a fair coin.

These standard tests show that corrected ΔEVI[7], though substantially smaller in magnitude than uncorrected, nonetheless shows a highly significant increase in forest greenness.

Second, we ask whether the smaller, but statistically significant, green up seen in the data from Morton *et al.* (Fig. 1) is biologically meaningful in terms of consistency with mechanisms and magnitude of seasonal changes in canopy-scale biophysics observed on the ground. We find that at an intensively measured site, significant dry-season increases in leaf area index are driven by coordinated flushing of new leaves, which have higher near-infrared reflectance (Fig. 2a) (mechanisms

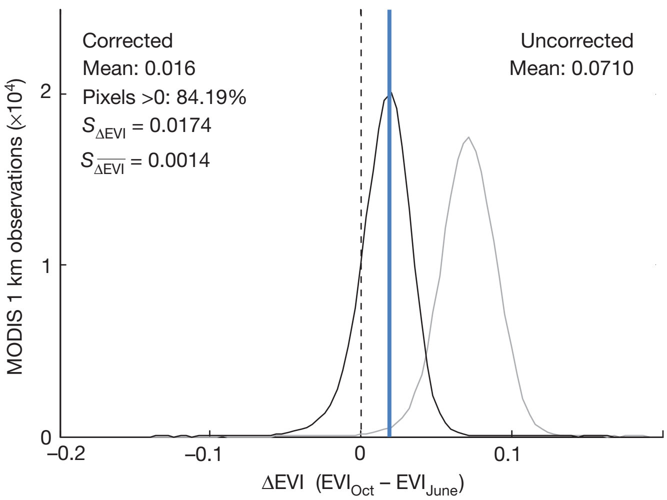


Figure 1 | Distribution of uncorrected and corrected ΔEVI. Reproducing figure 3b of ref. 7, with 95% confidence interval (shaded blue region), significantly excluding zero. We conservatively assume that only relatively large areas (1° × 1°, or ~10^4 MODIS pixels) are independent, giving 158 independent 1° × 1° patches that include valid pixels, and 95% confidence interval: $\overline{\Delta EVI} \pm Z S_{\overline{\Delta EVI}} = 0.016 \pm 0.0027$, where Z is 1.96 (the 95% Z-score), and $S_{\overline{\Delta EVI}}$ is the standard error of $\overline{\Delta EVI}$ (derived from ΔEVI standard deviation as $S_{\Delta EVI} / \sqrt{158}$). The probability of observing 84.2% positive values (or heads from fair coin flips) out of 158 observations is $P < 10^{-15}$ (binomial test).

that Morton *et al.*[7] hypothesized could drive true increases in satellite-observed EVI). Leaf flushing is followed, after 1 to 2 months, by increases in photosynthetic capacity derived from CO_2 fluxes measured at eddy flux towers (Fig. 2a). This correlation—1-month-lagged photosynthetic capacity with leaf area index, $r = +0.90$, and with MAIAC EVI, $r = +0.89$, where r is Pearson's correlation coefficient, and the time lag is for new leaves to develop their photosynthetic capacity[14]—establishes a link between eddy flux measurements and biophysical properties observable from satellites.

On the basis of this link, we find that increases in dry-season greenness seen by corrected EVI products (whether those of ref. 7 or the MAIAC EVI of Lyapustin *et al.*[10]; Fig. 2b) are real and consistently correlated with photosynthetic capacity increases seen at towers within the region analysed by Morton *et al.* (including adjustment for possible sun-angle effects on canopy illumination). This suggests that even the smaller corrected ΔEVI[7] reflects mechanisms of canopy changes actually observed on the ground, and is therefore biologically meaningful.

The analysis in Morton *et al.*[7] is, notably, stimulating a productive re-examination of the methodology, meaning and magnitude of remote sensing indices, their artefacts, and their relation to field studies on the ground[6,12]. However, we believe that the primary substantive finding of Morton *et al.* of consistent canopy structure and greenness is incorrect. Both satellite remote sensing and ground-based observations show dry-season increases in greenness and biophysical properties associated with canopy photosynthesis across scales, from individual leaves to ecosystems to regions, in support of the conclusion that Amazon forests green up with sunlight in the dry season[3,14].

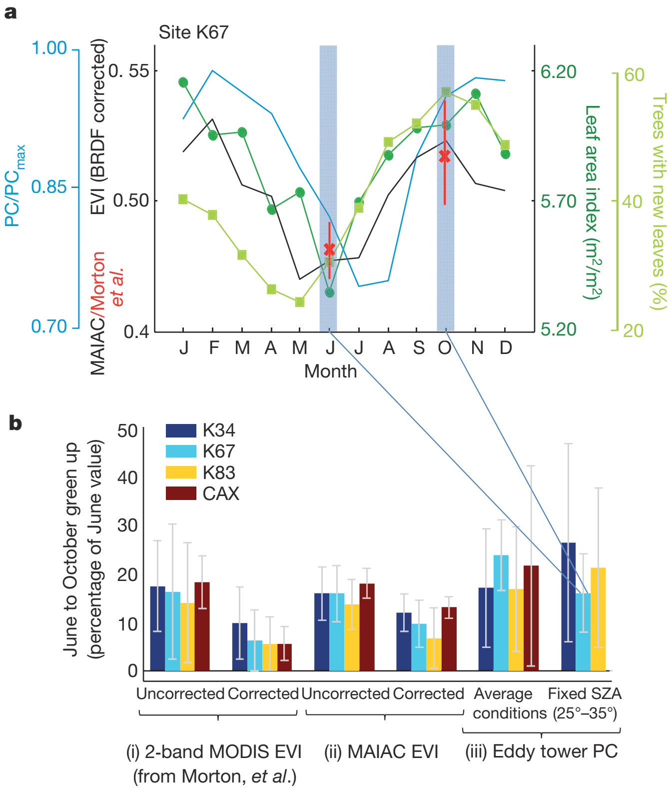


Figure 2 | Seasonality of vegetation metrics. a, Corrected EVI (from extended data figure 7 of ref. 7; red 'X' in June and October), and average cycle of corrected MAIAC EVI, a product that also corrects sun-sensor geometry (to nadir view, and 45° sun angle; black line)[10]; leaf area index (dark green)[15]; percentage of trees with new leaves (light green)[15]; and tower-derived photosynthetic capacity (PC/PC_{max}, blue line; see Methods), all from site K67 (ref. 4). BRDF, bidirectional reflectance distribution function. **b**, Green up at four tower sites for (i) EVI (corrected as in **a** and uncorrected); (ii) MAIAC EVI (corrected and uncorrected); and (iii) tower-derived photosynthetic capacity (with and without fixed solar zenith angle (SZA), showing potential effects of changing solar illumination). Sites: K34 (Manaus), K67 and K83 (Santarém), and CAX (Caxiuanã National Forest, near Belém (CAX had insufficient data for fixed SZA analysis))[4]. All uncertainties are 95% confidence intervals.

Methods

For basin-wide analysis, we analysed ΔEVI, corrected on a per-pixel basis, in 197,651 valid pixels (from figure 3 of ref. 7, data courtesy of D. C. Morton). For tower comparisons, we averaged valid pixels from both extended data figure 7 of ref. 7 (corrected with a simplified approach that retrieved ΔEVI for more area, $\sim 2 \times 10^6$ pixels, including around towers), and from the independent sun-sensor geometry corrected MAIAC EVI product[10], in 5 km grids around towers (11 km around CAX, to obtain sufficient data; Fig. 2b). Tower-based 95% confidences are boot-strapped June–October changes in photosynthetic capacity. Photosynthetic capacity is eddy-flux-derived gross primary productivity (from ref. 4) averaged under reference environmental conditions, an estimate of photosynthetic infrastructure independent of environment and (with additional binning) sun angle. Reference bins were: light ($1{,}350 \pm 200\,\mu$mol m^{-2} s^{-1}), vapour pressure deficit (980 ± 200 Pa), relative irradiance (observed/clear-sky expected $= 0.6 \pm 0.1$), and solar zenith angle (25°–35°).

Scott R. Saleska[1], Jin Wu[1], Kaiyu Guan[2], Alessandro C. Araujo[3], Alfredo Huete[4], Antonio D. Nobre[5] & Natalia Restrepo-Coupe[4]
[1]Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA.
email: saleska@email.arizona.edu
[2]Environmental and Earth System Science, Stanford University, Stanford, California 94305, USA.
[3]Embrapa Amazonia Oriental, Belem, Brazil.
[4]Plant Functional Biology and Climate Change Cluster, University of Technology Sydney, Sydney, New South Wales 2007, Australia.
[5]National Institute for Space Research (INPE) and National Institute for Amazonian Research (INPA), São José dos Campos, Brazil.

Received 21 November 2014; accepted 15 October 2015.

1. Borchert, R. Soil and stem water storage determine phenology and distribution of tropical dry forest trees. *Ecology* **75,** 1437–1449 (1994).
2. Wright, S. J. & Van Schaik, C. P. Light and the phenology of tropical trees. *Am. Nat.* **143,** 192–199 (1994).
3. Huete, A. R. *et al.* Amazon rainforests green-up with sunlight in dry season. *Geophys. Res. Lett.* **33,** L06405 (2006).
4. Restrepo-Coupe, N. *et al.* What drives the seasonality of productivity across the Amazon basin? A cross-site analysis of eddy flux tower measurements from the Brasil flux network. *Agric. For. Meteorol.* **182–183,** 128–144 (2013).
5. Shuttleworth, W. J. Evaporation from Amazonian rain forest. *Proc. R. Soc. Lond. B Biol. Sci.* **233,** 321–346 (1988).
6. Guan, K. *et al.* Photosynthetic seasonality of global tropical forests constrained by hydroclimate. *Nature Geosci.* **8,** 284–289 (2015).
7. Morton, D. C. *et al.* Amazon forests maintain consistent canopy structure and greenness during the dry season. *Nature* **506,** 221–224 (2014).
8. Jupp, D. L. B. & Strahler, A. H. A hotspot model for leaf canopies. *Remote Sens. Environ.* **38,** 193–210 (1991).
9. Myneni, R. *et al.* Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sens. Environ.* **83,** 214–231 (2002).
10. Lyapustin, A. I. *et al.* Multi-angle implementation of atmospheric correction for MODIS (MAIAC): 3. Atmospheric correction. *Remote Sens. Environ.* **127,** 385–393 (2012).
11. Huete, A. *et al.* Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **83,** 195–213 (2002).
12. Bi, J. *et al.* Sunlight mediated seasonality in canopy structure and photosynthetic activity of Amazonian rainforests. *Environ. Res. Lett.* **10,** (2015).
13. Sheskin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures* 4th edn (Chapman & Hall/CRC, 2007).
14. Wu, J. *et al.* Leaf development and demography explain photosynthetic seasonality in Amazon evergreen forests. *Science* **351,** 972–976 (2016).
15. Brando, P. M. *et al.* Seasonal and interannual variability of climate and vegetation indices across the Amazon. *Proc. Natl Acad. Sci. USA* **107,** 14685–14690 (2010).

Author Contributions S.R.S. designed the statistical analysis and wrote the initial draft of the paper; J.W. implemented the statistical analysis; K.G. provided MAIAC EVI data and analysis; A.C.A. contributed eddy flux data; A.H. contributed analysed EVI data; A.D.N. contributed eddy flux data and insights; and N.R.C. integrated the multi-tower datasets. All authors contributed to writing the final version.

Competing Financial Interests Declared none.

doi:10.1038/nature16457

Morton *et al.* reply

REPLYING TO S. R. Saleska *et al. Nature* **531,** http://dx.doi.org/10.1038/nature16457 (2016)

Multiple mechanisms could lead to upregulation of dry-season photosynthesis in Amazon forests, including canopy phenology and illumination geometry. We specifically tested two mechanisms for phenology-driven changes in Amazon forests during dry-season months, and the combined evidence from passive optical and lidar satellite data[1] was incompatible with large net changes in canopy leaf area or leaf reflectance suggested by previous studies[2–5]. We therefore hypothesized[1] that seasonal changes in the fraction of sunlit and shaded canopies, one aspect of bidirectional reflectance effects in Moderate Resolution Imaging Spectroradiometer (MODIS) data, could alter light availability for dry-season photosynthesis and the photosynthetic capacity of Amazon forests without large net changes in canopy composition. Subsequent work supports the hypothesis that seasonal changes in illumination geometry and diffuse light regulate light saturation in Amazon forests[6,7]. These studies clarify the physical mechanisms that govern light availability in Amazon forests from seasonal variability in direct and diffuse illumination. Previously, in the debate over light limitation of Amazon forest productivity, seasonal changes in the distribution of light within complex Amazon forest canopies were confounded with dry-season increases in total incoming photosynthetically active radiation[2,3,8]. In the accompanying Comment[9], Saleska *et al.* do not fully account for this confounding effect of forest structure on photosynthetic capacity.

Saleska *et al.*[9] investigated one of the three lines of evidence in our paper to argue that near-zero seasonal changes in corrected MODIS enhanced vegetation index (EVI) are actually non-zero (figure 1 in ref. 9; 0.071 to 0.016, a 77% reduction). Following this logic, our data also show a small but statistically significant decrease in normalized difference vegetation index (NDVI; extended data figure 4 in ref. 1), a pattern that we attributed to residual artefacts from changes in sun-sensor geometry, as no leaf-level mechanism for increased forest productivity generates opposing responses in these vegetation indices (see supplementary discussion in ref. 1). Indeed, the comparison between NDVI and EVI responses is a useful diagnostic tool[1] that could have been used to investigate residual bidirectional reflectance effects in multiangle implementation of atmospheric correction (MAIAC) data (figure 2 in ref. 9).

In isolation, MODIS data provide limited insight into the mechanisms for seasonal changes in Amazon forests[9]. MODIS EVI is primarily sensitive to changes in near-infrared reflectance[1,4,10], not photosynthetically active radiation absorption that drives forest productivity. Saleska *et al.* misrepresent data from extended data figure 7 of ref. 1 as fully corrected in their figure 2 (ref. 9), and further confound seasonal changes through spatial averaging of 1 km^2 data over large regions (25–121 km^2). A previous study using 1 km^2 data for these same tower sites shows little or no seasonality in MAIAC EVI[11] (see supplementary figure 5 in ref. 1).

One of the key messages from our study was the need for careful attention to uncertainty in satellite-based measurements of forest seasonality. The presentation of *in situ* and satellite data by Saleska *et al.* (figure 2a in ref. 9), and the MAIAC product in general, could be improved with quantitative estimates of uncertainty to support assertions of forest seasonality.

Subtle variability in canopy structure and reflectance properties of Amazon forests remains a key area for further study, particularly with large-scale field studies[12], to better understand the spatial and temporal heterogeneity of leaf phenology strategies in Amazonia[13]. Other mechanisms for seasonal changes in photosynthetic capacity also merit further investigation, including how diurnal and seasonal variability in illumination alter the distribution of photosynthetically active radiation at the leaf level[1,6,7,14]. NASA satellite data remain an important foundation for future research on tropical forest dynamics, within the limits of calibration, measurement, and model uncertainty that can be realistically achieved with space-based sensors.

Douglas C. Morton[1], Jyoteshwar Nagol[2,3], Claudia C. Carabajal[1,4], Jacqueline Rosette[1,2,5], Michael Palace[6], Bruce D. Cook[1], Eric F. Vermote[1], David J. Harding[1] & Peter R. J. North[5]
[1]NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.
email: douglas.morton@nasa.gov
[2]University of Maryland, College Park, Department of Geographical Sciences, College Park, Maryland 20742, USA.
[3]Global Land Cover Facility, College Park, Maryland 20740, USA.
[4]Sigma Space Corporation, Lanham, Maryland 20706, USA.
[5]Swansea University, Department of Geography, Singleton Park, Swansea SA2 8PP, UK.
[6]Earth Systems Research Center, University of New Hampshire, Durham, New Hampshire 03824, USA.

1. Morton, D. C. *et al.* Amazon forests maintain consistent canopy structure and greenness during the dry season. *Nature* **506,** 221–224 (2014).
2. Huete, A. R. *et al.* Amazon rainforests green-up with sunlight in dry season. *Geophys. Res. Lett.* **33,** L06405 (2006).
3. Brando, P. M. *et al.* Seasonal and interannual variability of climate and vegetation indices across the Amazon. *Proc. Natl Acad. Sci. USA* **107,** 14685–14690 (2010).
4. Samanta, A. *et al.* Seasonal changes in leaf area of Amazon forests from leaf flushing and abscission. *J. Geophys. Res.* **117,** G01015 (2012).
5. Myneni, R. B. *et al.* Large seasonal swings in leaf area of Amazon rainforests. *Proc. Natl Acad. Sci. USA* **104,** 4820–4823 (2007).
6. Morton, D. C. *et al.* Amazon forest structure generates diurnal and seasonal variability in light utilization. *Biogeosciences Discuss.* **12,** 19043–19072 (2015).
7. Rap, A. *et al.* Fires increase Amazon forest productivity through increases in diffuse radiation. *Geophys. Res. Lett.* **42,** 4654–4662 (2015).
8. Restrepo-Coupe, N. *et al.* What drives the seasonality of productivity across the Amazon basin? A cross-site analysis of eddy flux tower measurements from the Brasil flux network. *Agric. For. Meteorol.* **182–183,** 128–144 (2013).
9. Saleska, S. R. *et al.* Dry-season greening of Amazon forests. *Nature* **531,** http://dx.doi.org/10.1038/nature16457 (2016).
10. Galvão, L. S. *et al.* On intra-annual EVI variability in the dry season of tropical forest: a case study with MODIS and hyperspectral data. *Remote Sens. Environ.* **115,** 2350–2359 (2011).
11. Guan, K. *et al.* Photosynthetic seasonality of global tropical forests constrained by hydroclimate. *Nature Geosci.* **8,** 284–289 (2015).
12. Soudani, K. & Francois, C. Remote sensing: a green illusion. *Nature* **506,** 165–166 (2014).
13. Chave, J. *et al.* Regional and seasonal patterns of litterfall in tropical South America. *Biogeosciences* **7,** 43–55 (2010).
14. Alton, P. B. *et al.* The impact of diffuse sunlight on canopy light-use efficiency, gross photosynthetic product and net ecosystem exchange in three forest biomes. *Glob. Change Biol.* **13,** 776–787 (2007).

doi:10.1038/nature16458

For News & Views online, go to
nature.com/newsandviews

MATERIALS SCIENCE

How crystals get an edge

Microscopy reveals how nanowires of a widely used semiconductor grow during preparation. The findings will allow the crystal phases of such nanowires to be engineered — a crucial advance for materials science. SEE ARTICLE P.317

ANNA FONTCUBERTA I MORRAL

The twinkling facets of single crystals in mineral collections have inspired breathtaking artwork and magnificent buildings, such as the Sagrada Família in the Catalan city of Barcelona, Spain. Yet our understanding of the nucleation (the formation of tiny crystal seeds at the start of crystallization) and growth of even the smallest filamentary crystals, including nanowires, is fragmentary. This slows the implementation of nanowire-based technology in many areas of science, including electronics, photonics and quantum information. On page 317 of this issue, Jacobsson and colleagues[1] show how to control the crystal form adopted by nanowires at nucleation, thereby allowing the structure to be tailored for particular functions, and solving one of the most fundamental challenges in materials science and engineering.

Polytypism is the ability of a compound to exist as various crystal forms that differ only in terms of the sequence in which 'bilayer structures' of atoms stack in a given direction within a crystal lattice. Bilayer structures are arrangements of atom pairs connected by bonds, and can be thought of as subunits from which all the possible crystal types of a compound can be constructed. An archetype of polytypism in bulk crystals is that of silicon carbide (SiC), for which three bilayer structures (defined as A, B and C) are needed. Silicon carbide has more than 200 polytypes, corresponding to different combinations of the bilayers[2]. For cubic crystal systems, including the cubic structures adopted by silicon carbide and gallium arsenide (GaAs, a widely used semiconductor), the most common polytypes correspond to the zinc-blende (ABCABC) and wurtzite (ABAB) crystal structures.

The polytype formed is determined at the nanometre scale during nucleation. At this scale, the high surface-to-volume ratio of crystals can stabilize polytypes and specific crystal phases that are unstable in the bulk form. The formation of polytypes in nanowires is especially intriguing. Nanowires are often made using the vapour–liquid–solid (VLS) method, in which a liquid-metal droplet acts as a catalyst that gathers gaseous precursors and transforms them into a solid that grows beneath the droplet (Fig. 1). The growth conditions determine which polytypes form for a given diameter of the nanowire.

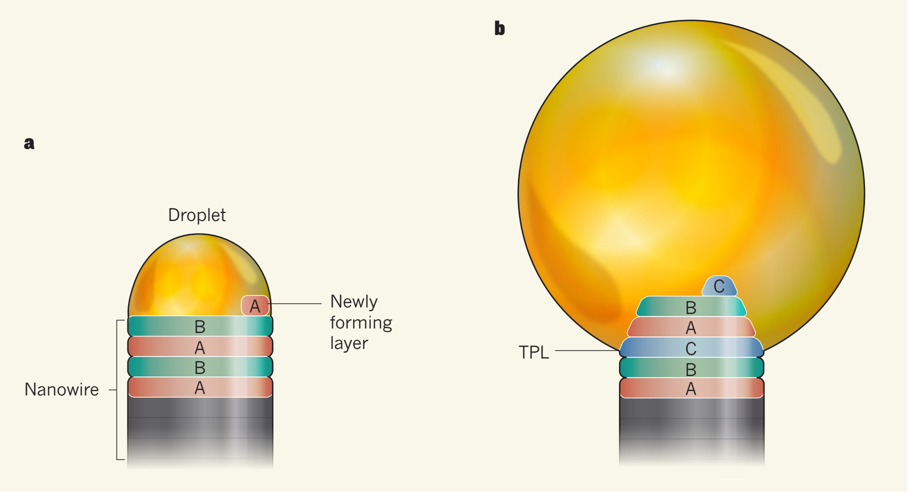


Figure 1 | Nanowire growth. In the vapour–liquid–solid method, a liquid-metal droplet acts as a catalyst that transforms gaseous precursors (not shown) into a crystalline solid that grows as a nanowire beneath the droplet. Different crystal phases can grow, consisting of different sequences of layers that contain distinct structural arrangements of atoms; layer types are denoted A, B and C. **a**, Jacobsson *et al.*[1] observed that, when the contact angle of the droplet to the nanowire is close to 90°, the liquid–solid interface is flat, and favours ABAB stacking (the wurtzite crystal phase). **b**, Contact angles larger or smaller than 90° result in the formation of a crystal edge that starts at the triple-phase line (TPL, which separates the vapour, liquid and solid phases), and favours the ABCABC stacking of the zinc-blende phase.

Several theories[3–5] have addressed the question of which polytype forms during VLS nanowire formation, and have proposed that the triple-phase line (TPL, also known as the trijunction) — which separates the vapour, liquid and solid phases in VLS processes — has a decisive role. According to these theories, nucleation at the TPL favours formation of a wurtzite phase, whereas nucleation outside the TPL favours the zinc-blende phase. Control of whether nucleation occurs at or outside the TPL might be achieved by modifying the contact angle between the VLS droplet and the nanowire, but explicit experimental evidence to support this hypothesis has been lacking. This was particularly true for the technologically interesting family of arsenide and phosphide semiconductors.

Jacobsson *et al.* provide the first experimental evidence of the role of the TPL and the contact angle in polytype formation, going beyond simple growth conditions. By watching GaAs nanowires grow using a transmission electron microscope, they have determined the parameters that lead to wurtzite or zinc-blende structures. More specifically, they observed the changes of crystal phase that occur alongside variations in the flux of gaseous GaAs precursors (trimethylgallium and arsine).

The authors began by observing nanowire growth using VLS conditions that lead to formation of the wurtzite phase. They found that each new bilayer formed at the TPL, creating a complete, flat layer across the top of the crystal (Fig. 1a). The liquid droplet became larger and richer in gallium as the flux of arsine was reduced.

When the droplet reached a certain size, Jacobsson and colleagues observed the sudden appearance of an edge — a facet of the nanowire that starts at the TPL within the liquid–solid interface (Fig. 1b). The growth dynamics of the nanowire then changed drastically. New bilayers formed much faster than before, and each was accompanied by oscillation of the edge — that is, the position of the edge moved

outwards from the bilayer and then jumped back again. Once an edge appeared, further growth on the nanowire quickly adopted the zinc-blende structure. Previously grown wurtzite segments partially dissolved at the edge, but regrew as wurtzite.

Jacobsson *et al.* propose that the size of the droplet determines both its contact angle with the nanowire and the morphology of the liquid–solid interface. They also suggest that contact angles of about 90° are compatible with a flat liquid–solid interface and with nucleation of bilayers at the TPL. By contrast, smaller or larger contact angles lead to the formation of an edge at the liquid–solid interface, starting at the TPL — which suppresses nucleation of new bilayers at the TPL, enabling the zinc-blende structure to form.

These results extend our knowledge far beyond the orthodox understanding of polytype formation, and provide a path towards crystal-phase design: it should now be possible to engineer VLS systems to select the crystal phase of GaAs that forms. But can polytypism be tailored for different VLS catalysts, or even in their absence? The answer could be crucial to avoiding the unintentional incorporation of impurities from catalysts into nanowires.

The functional behaviour of classical semiconductor heterostructures[6] (combinations of two different semiconductors) is controlled by the change of composition that occurs at the interface between the component materials — the interface modifies the potential-energy landscape of charge carriers such as negatively charged electrons and positively charged 'holes'. In his physics Nobel lecture, Herbert Kroemer introduced the concept of engineering heterostructures to 'teach electrons new tricks' in electronic and optoelectronic applications[6]. The unintentional formation of different polytypes in heterostructures is detrimental to such applications. But Jacobsson and co-workers' findings about how to make polytypes reproducibly mean that we can add crystal-phase engineering to the list of techniques by which we teach new tricks to electrons, holes and their associated spins[7], and might enable the development of new semiconductor applications. ■

Anna Fontcuberta i Morral *is in the Laboratory of Semiconductor Materials, Institute of Materials, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland.*
e-mail: anna.fontcuberta-morral@epfl.ch

1. Jacobsson, D. *et al. Nature* **531,** 317–322 (2016).
2. Bechstedt, F. *et al. Phys. Stat. Sol. B* **202,** 35–62 (1997).
3. Glas, F., Harmand, J.-C. & Patriarche, G. *Phys. Rev. Lett.* **99,** 146101 (2007).
4. Krogstrup, P. *et al. Phys. Rev. Lett.* **106,** 125505 (2011).
5. Dubrovskii, V. G., Sibirev, N. V., Harmand, J. C. & Glas, F. *Phys. Rev. B* **78,** 235301 (2008).
6. Kroemer, H. *Rev. Mod. Phys.* **73,** 783 (2001).
7. Falk, A. L. *et al. Nature Commun.* **4,** 1819 (2013).

BIOMEDICINE

Visionary stem-cell therapies

Stem-cell engineering has allowed successful cornea transplantations in rabbits and the regeneration of transparent lens tissue in children, demonstrating the therapeutic potential of this approach. SEE ARTICLE P.323 & LETTER P.376

JULIE T. DANIELS

To correctly refract light onto the retina at the back of the eye, the cornea and lens must remain transparent throughout life. Treatments for opacification of the cornea or lens involve donor transplants or artificial implants, respectively, but these procedures can be risky. Alternative strategies for treating such ocular disorders would be to transplant tissue grown in the laboratory from stem cells, or to coax resident stem cells to regenerate normal tissue in the body. Two papers in this issue, by Hayashi *et al.*[1] (page 376) and Lin *et al.*[2] (page 323), report work that advances these possibilities.

Blindness represents a great clinical and economic burden worldwide. Corneal transplantation has been the gold standard for restoring transparency since the first successful transplant in 1905 (ref. 3). However, despite the fact that corneal transplants are less likely to induce an immune response than transplants to other sites in the body, grafts can be rejected by the host body within five years[4].

One promising strategy for avoiding rejection involves tissue engineering using the patient's own (autologous) cells. This approach has proved successful in treating people with ocular chemical burns, in whom autologous cells called limbal stem cells were used to replace the epithelial cells that make up the outermost layer of the cornea[5,6]. However, it is not always feasible to harvest sufficient stem cells to grow in the laboratory. As an alternative, the conversion of adult cells into induced pluripotent stem cells (iPSCs), which can develop into any cell type[7], could supply enough cells for therapy.

During embryonic development, ocular tissue is formed from three tissue layers, one of which, the surface ectoderm, gives rise to the corneal epithelium and lens. Hayashi *et al.*[1] grew human iPSCs *in vitro* under conditions that promoted the creation of a structure

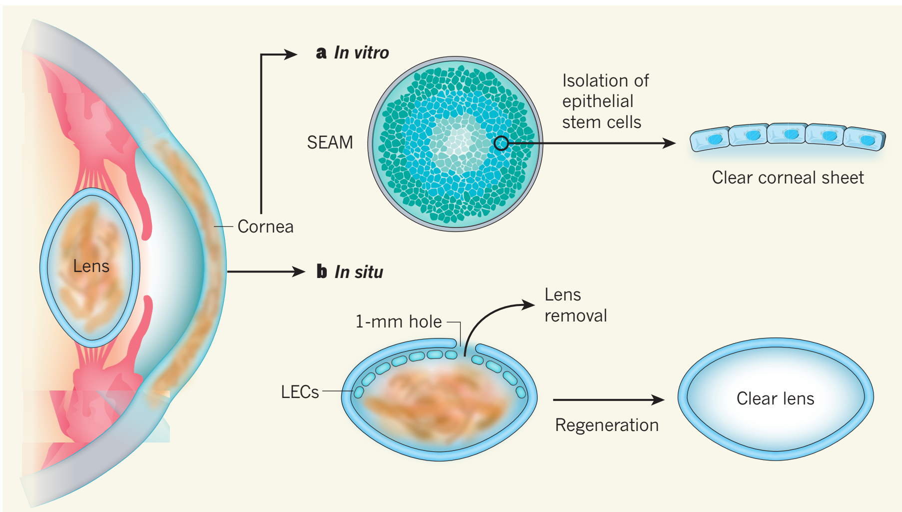


Figure 1 | Sight for sore eyes. Opacity in the cornea or the lens of the eye can cause blindness, and current treatment strategies are less than ideal. **a,** To produce corneal sheets that could be transplanted into rabbits, Hayashi *et al.*[1] cultured induced pluripotent stem cells (iPSCs, which can give rise to all body cell types) in conditions that promoted the development of a SEAM — a structure containing four distinct zones and resembling the embryonic eye. Isolation and further culture of epithelial stem cells from zone 3 of the SEAM successfully produced corneal sheets for transplantation. **b,** Opaque lenses are normally removed and replaced with artificial implants. Lin *et al.*[2] devised a minimally invasive method for lens removal, which left lens epithelial stem cells (LECs) intact. These LECs regenerated the lens in rabbits, macaques and children, eliminating the need for implants.

that they dubbed the self-formed ectodermal autonomous multi-zone (SEAM), which contained four defined concentric zones that in some ways mimicked the developing eye. The authors found that different SEAM zones contained cells with characteristics of the ocular surface ectoderm, the lens, the neuro-retina and the retinal pigment epithelium.

Blocking BMP signalling — an intracellular pathway required for the development of surface ectoderm cells — abolished zone 3 of the SEAM. Hayashi and colleagues tested the therapeutic potential of cells from this zone by harvesting the cells and selecting those that expressed genes characteristic of epithelial stem cells (Fig. 1a). The authors cultured transplantable sheets of corneal epithelium from the selected cells, and demonstrated that these sheets could restore a healthy corneal epithelium in rabbits in which corneal epithelial stem cells had been experimentally depleted.

Cataracts, which cause sight-threatening lens clouding[8], are surgically treated by removing the lens from its supporting capsule and replacing it with an artificial intraocular lens (IOL). In children with congenital cataracts, a major cause of childhood blindness, the success of IOL implantation is limited[9,10] — surgery can cause opacity in the line of vision and, because the eye is still growing, it is difficult to provide good vision with spectacles. Rather than attempting to create a living lens *in vitro*, Lin *et al.*[2] investigated the possibility of regenerating a naturally transparent lens in the body.

The authors discovered that lens epithelial stem/progenitor cells (LECs) expressing the genes *PAX6* and *SOX2* self-renew and differentiate into lens fibre cells that can form a 3D transparent lens-like structure that refracts light. In mice, mutation of the stem-cell-maintenance gene *Bmi-1*, which is expressed in LECs, impaired LEC proliferation and induced cataract formation. These data led Lin *et al.* to reason that, by refining the technique for surgical cataract removal to minimize the damage done to LECs *in situ*, they could promote lens regeneration (Fig. 1b).

In rabbits, the authors' minimally invasive technique led to lens regeneration around seven weeks after surgery. The approach achieved similar results in macaques, in which lens regeneration took several months and no complications arose. Finally, the researchers performed a clinical trial, in which transparent lenses were regenerated in both eyes of 12 infants within 3 months, all without complication.

These two studies illustrate the remarkable regenerative and therapeutic potential of stem cells. Hayashi and colleagues' approach involved substantial *in vitro* cell manipulation to obtain a sheet of cultured corneal epithelium for transplantation. When considering the expense involved in following good manufacturing practices for cell therapies, the current protocol is unlikely to be economically viable. However, the real value of this research lies in the possibility that the SEAM model will facilitate the discovery of fundamental mechanisms that underlie the early development of each type of ocular tissue. Such an understanding might eventually enable *in situ* manipulation of stem-cell populations throughout the eye, as Lin *et al.* have elegantly shown to be achievable for the lens. Furthermore, lens regeneration might also turn out to be possible in ageing adults in whom LEC proliferation has declined — for example, research on the SEAM could identify small molecules able to stimulate such regeneration.

Whether either of the reported therapies will lead to cornea or lens transparency that can be maintained in the long term remains uncertain. However, these exciting studies take us away from simple therapies that involve like-for-like replacement of single mature cell types, and open up the possibility of therapeutic manipulation of the broader stem-cell environment in the eye. ■

Julie T. Daniels *is at the University College London Institute of Ophthalmology, London EC1V 9EL, UK.*
e-mail: j.daniels@ucl.ac.uk

1. Hayashi, R. *et al. Nature* **531,** 376–380 (2016).
2. Lin, H. *et al. Nature* **531,** 323–328 (2016).
3. Armitage, W. J., Tullo, A. B. & Larkin, D. F. P. *Brit. J. Ophthalmol.* **90,** 1222–1223 (2006).
4. Borderie, V. M. *et al. Ophthalmology* **116,** 2354–2360 (2009).
5. Pellegrini, G. *et al. Lancet* **349,** 990–993 (1997).
6. Rama, P. *et al. N. Engl. J. Med.* **363,** 147–155 (2010).
7. Takahashi, K. & Yamanaka, S. *Cell* **126,** 663–676 (2006).
8. Stevens, G. A. *et al. Ophthalmology* **120,** 2377–2384 (2013).
9. Visser, N., Bauer, N. J. C. & Nuijts, R. M. M. A. *J. Cataract Refract. Surg.* **39,** 624–637 (2013).
10. Mamalis, N., Davis, B., Nilson, C. D., Hickman, M. S. & LeBoyer, R. M. *J. Cataract Refract. Surg.* **30,** 2209–2218 (2004).

This article was published online on 9 March 2016.

GLOBAL WARMING

China's contribution to climate change

Carbon dioxide emissions from fossil-fuel use in China have grown dramatically in the past few decades, yet it emerges that the country's relative contribution to global climate change has remained surprisingly constant. **See Letter p.357**

DOMINICK V. SPRACKLEN

In December 2015, world leaders agreed to limit the increase in global average temperature to less than 2 °C above pre-industrial temperatures (see *Nature* **528,** 315–316; 2015). Meeting this aspiration will require large and rapid reductions in greenhouse-gas emissions, making it imperative to understand and account for the emissions from different countries. China has undergone rapid economic development over the past few decades and now has one of the world's largest economies — and greenhouse-gas emissions to match. On page 357 of this issue, Li *et al.*[1] comprehensively assess China's contribution to climate change and explore how this has altered as the Chinese economy has grown.

Humans affect Earth's climate through many mechanisms by changing the abundance of greenhouse gases and air pollutants, and by altering the reflectivity of Earth's surface through changes in land use. The relative strengths of these different drivers can be compared through a metric known as radiative forcing, which quantifies the impact of each process on Earth's energy budget.

Li *et al.* used a model that couples biogeochemistry and climate to estimate China's contribution to global radiative forcing over the period 1980–2010. Crucially, they account for almost all anthropogenic drivers of climate change. They find that China's relative contribution to global radiative forcing from carbon dioxide emissions associated with fossil-fuel use increased almost threefold in these 30 years. This is to be expected, given the surge in China's economy over this period. More surprisingly, they find that China's relative contribution to total global radiative forcing has remained at 10% over this time.

To understand the reasons behind this remarkable result, Li and colleagues made a detailed analysis of the different drivers of radiative forcing. They found that the air pollutants that cause China's notorious pollution haze have had complex effects on climate, counteracting some of the increase in radiative forcing from greenhouse gases. Some components of air pollution, such as black-carbon particles, absorb sunlight and warm Earth's climate. By contrast, sulfate particles scatter light, resulting in climate cooling.

Over the past few decades, China's relative contribution to global radiative forcing from sulfate has increased dramatically. This

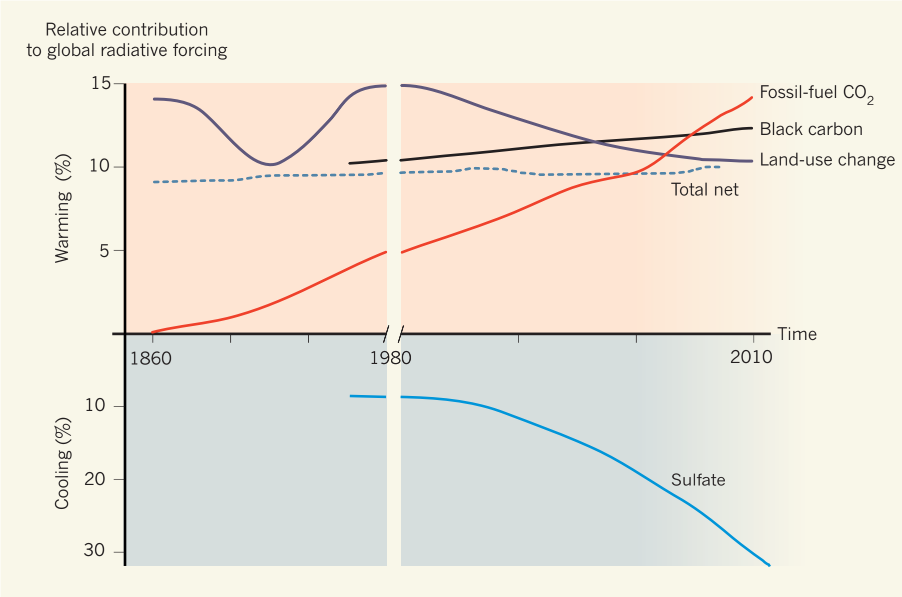


Figure 1 | No net change. Li *et al.*[1] report that China's relative contribution to global radiative forcing — a measure of how strongly different factors affect climate change — has remained constant over the past three decades (broken blue line). The total net effect has several contributors. Carbon dioxide emissions from fossil-fuel combustion have increased, as have black-carbon emissions, both of which lead to climate warming (positive radiative forcing). Land-use change can also contribute to warming, but the effects of this have declined. Conversely, sulfate emissions that cool climate have increased, and the negative radiative forcing associated with this has offset that from warming factors. Plotted lines are approximate.

is because Chinese sulfate emissions soared at the same time that Europe and the United States instigated controls that slashed their sulfate emissions. It has long been known that some air pollutants cool the climate[2]; what is remarkable in the present study is that the concurrent changes in different emissions have led to a stable overall contribution of China to global radiative forcing (Fig. 1).

Air pollution is a serious environmental issue in China, where 1.3 million people die each year because of exposure to poor-quality air outdoors[3]. Reductions in the emissions of air pollutants are urgently required to improve air quality, but this will also affect Earth's climate. Li *et al.* find that the current composition of Chinese air pollution causes almost no net radiative forcing — the cooling effects of sulfate aerosols balance the warming impacts of black-carbon emissions.

This means that it will be difficult to achieve rapid reductions in near-term global warming through the control of Chinese air pollutants overall — a focus on greenhouse-gas emissions in particular will be required. It also means that carefully managed mitigation of air pollution that focuses on reducing both black-carbon and sulfate emissions might have a minimal impact on climate, because their effects seem to counteract each other. Controlling the combustion of solid fuels for cooking and heating in the home is important in this context, because domestic solid fuel accounts for 40% of Chinese black-carbon emissions[4] and causes half a million deaths annually through poor outdoor air quality[3,4].

Li and co-workers went on to explore China's contributions to emissions of CO_2 and methane from pre-industrial times (1750) to the present day. They find that China's relative contribution to radiative forcing from these greenhouse gases has remained remarkably constant over this much longer period as well. The extensive conversion of China's natural forests to agricultural land resulted in substantial CO_2 emissions in the early part of this period. The rate of deforestation has declined in recent decades, but this has been counteracted by increasing fossil-fuel emissions. China is now planting forests on a larger scale than anywhere else on the planet. These plantations sequester CO_2 from the atmosphere, so that Chinese forests are now a net sink of this gas.

Mitigating climate change and air quality without unintended consequences will require an understanding of many complex interactions. Current models, including the one used by Li *et al.*, do not cover many of these complexities. In particular, the authors' study does not consider the formation of secondary organic aerosols — which might dominate in the haze over China[5] — from gaseous pollutants. Detailed monitoring of Chinese air pollution is urgently needed to inform the development of effective mitigation policies[6].

Air pollutants also interact in complex ways with ecosystems: land-use change alters air quality[7], and deposition of pollution can alter forest growth and carbon sequestration[8]. But these effects are not included in many models. Recent work[9] has shown that fast-growing forest plantations in Europe store less biomass

50 Years Ago

The winds of change are sometimes almost indistinguishable from placid summer breezes. The decision of the British Government that British money must now be decimal … would bring some benefit, not disaster … There will be those who claim that the duodecimal system is better because twelve has several integral factors, though it is at least as sensible to argue that the base of all arithmetic should be a prime number in order that people should not be encouraged to manipulate vulgar fractions … More distantly, other feats of rationalization may now be attempted. Why not, for example, decimalize the day? … A decimalized day should be a much more practical proposition. From midnight to midnight would be a million new seconds. One per cent of a day, or 10^4 new seconds, would be a convenient sub-unit roughly equal to a quarter of what is now an hour. Astronomers and airline travellers alike would welcome — in due course — that further proof that decimals are not merely reasonable but inevitable.

From *Nature* 19 March 1966

100 Years Ago

Scientific men in their most august society are banded together "for the improvement of natural knowledge." They are by implication a body of students working in the temple of Nature for truth's sake alone, heedless of the world and its rewards. What they garner is their gift to the world: they fill another page in the Revelation that brings men nearer to the angels. Let a man wander into the world with his science as wares to sell for money profit, and he has passed from true brotherhood. Surely this idea, perhaps rather fancifully stated, is at the bottom of much of our exclusiveness.

From *Nature* 16 March 1916

and absorb more sunlight than do natural forests; both of these features reduce the forests' benefit to the climate. Whether similar issues are at play across China requires investigation, but it is possible that a greater focus than at present on protecting and restoring natural forests in China might also provide greater benefits for global climate. ■

Dominick V. Spracklen *is in the School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK.*
e-mail: dominick@env.leeds.ac.uk

1. Li, B. *et al. Nature* **531,** 357–361 (2016).
2. Fiore, A. M. *et al. Chem. Soc. Rev.* **41,** 6663–6683 (2012).
3. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. *Nature* **525,** 367–371 (2015).
4. Butt, E. W. *et al. Atmos. Chem. Phys.* **16,** 873–905 (2016).
5. Huang, R.-J. *et al. Nature* **514,** 218–222 (2014).
6. Kulmala, M. *Nature* **526,** 497–499 (2015).
7. Heald, C. L. & Spracklen, D. V. *Chem. Rev.* **115,** 4476-4496 (2015).
8. Mahowald, N. *Science* **334,** 794–796 (2011).
9. Naudts, K. *et al. Science* **351,** 597–600 (2016).

MICROROBOTICS

Swimmers by design

Scientists have created soft microrobots whose body shapes can be controlled by structured light, and which self-propel by means of travelling-wave body deformations similar to those exhibited by swimming protozoa.

IGOR S. ARANSON

Swarms of 'smart' microrobots scouting the human body, delivering medicines or assembling complex micromachines have long been a popular theme in blockbuster films and best-selling novels. Take, for example, the 1966 film *Fantastic Voyage*, or Michael Crichton's 2002 novel *Prey*. Although at present the concept remains in the realm of science fiction, researchers are taking strides towards bringing the vision closer to reality. Writing in *Nature Materials*, Palagi and colleagues[1] report a big advance towards this goal: the creation of a synthetic light-powered microrobot inspired by swimming *Paramecium* protozoa.

Designing a robust microscopic robotic swimmer that can navigate complex environments and perform useful functions is a key component of the quest. To operate autonomously or on demand, a microswimmer should be able to harvest energy, propel itself through fluid towards its target and respond to external signals. Energy is needed both to overcome the friction of the fluid and to maintain motion for a long time — up to an hour for some biomedical applications.

Several designs exist. One is a microscopic gold–platinum rod that self-propels in an aqueous solution of hydrogen peroxide (H_2O_2; ref. 2). The microrod decomposes H_2O_2 and transfers the energy released by this process to the water. Ensuing water flow pushes the rod forward like a miniature submarine. A similar design[3,4] sees a microscopic, rocket-shaped, composite platinum–polymer tube generate gas bubbles from the decomposition of H_2O_2, with the detaching bubbles propelling the microrocket.

External magnetic and electric fields can also be used to power microswimmers[5–9]. Some swimmers move like snakes or worms, gliding on the surface of water by periodically bending their bodies[8,9]. Scientists foresee microswimmers being used to unclog arteries[10], and to deliver immobile sperm to fertilize eggs[11].

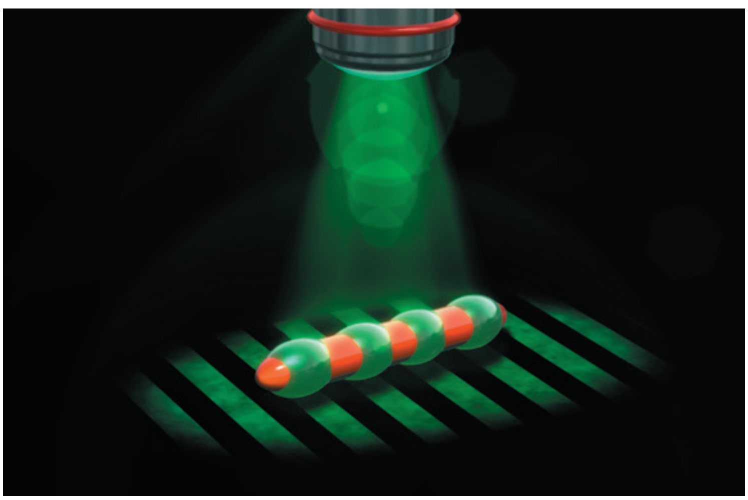
Nature has mastered highly effective means of micrometre-scale propulsion, exemplified by the rotation of helical bacterial flagella, and the wavy beating of the cilia (tiny hair-like structures) that cover *Paramecium*. This metachronal wave — the sequential movement of thousands of cilia — enables paramecia to swim at astounding speeds[12], up to ten body lengths per second. (For comparison, a dolphin barely makes two to three body lengths per second when in a hurry.)

Synthetic microswimmers exhibiting this amazing wavy propulsion would be highly desirable. However, implementing the coordinated wavy deformation of the microscopic swimmer's body presents a technical challenge: the need to manufacture myriad minuscule actuators and joints that can be individually controlled.

Palagi and colleagues propose an elegant means of microscopic propulsion afforded by travelling waves. Instead of a cumbersome array of addressable actuators, the authors use a synthetic polymer — a liquid-crystal elastomer. This rubbery material, made up of molecules oriented in a certain direction, exhibits remarkably strong coupling between molecule orientation and mechanical deformation. The liquid-crystal elastomer elongates when the molecules are fully aligned with each other and shrinks when this molecular ordering is lost, typically when it is heated or exposed to intense light. As a result, this material can be highly sensitive to external stimuli such as light and heat[13].

To make a swimmer, the authors illuminated a millimetre-long rod of this photoresponsive material with a laser beam, using a rectangular array of microscopic computer-controlled mirrors to project a moving sequence of light and dark stripes onto the rod (see Fig. 1). The material responded by expanding or shrinking. Thus, a moving sequence of projected light and dark stripes created a pattern of travelling bumps along the rod, resembling the wavy beating of cilia. Remarkably, the rod swam.

Palagi *et al.* could tune the speed of their soft microrobots by adjusting the speed at which the projected stripes moved, and the motion could be started or stopped by turning the light on or off, respectively. By changing the light patterns, it was also possible to control several microswimmers simultaneously and coerce them into rotating or travelling along a designated path. Theoretically, their swimming speed should increase proportionally with the speed at which the light stripes move. However, the authors found that the material's time response limits the



ALEJANDRO POSADA/MAX PLANCK INST. INTELLIGENT SYSTEMS

Figure 1 | Light-powered swimmer. A microscope (top) projects a moving sequence of light (green) and dark (black) stripes onto a photoresponsive, millimetre-long polymer rod, inducing periodic deformations (visible as bumps) on the rod's surface. Palagi *et al.*[1] show that these travelling deformations propel the soft rod through a fluid in a manner that mimics the locomotion of protozoa.

maximum speed to about 40 micrometres per second (or, in terms of body length, about 30 times slower than a dolphin).

This work demonstrates a first step towards truly bioinspired self-propulsion. However, any practical applications would require optical access. To make these robots more competitive with other synthetic swimmers, their size needs to be reduced to the micrometre scale and their relative swimming speed greatly increased. There are no serious technical limitations on manufacturing smaller and smaller polymer rods, but increasing the swimming speed is a different issue that comes down to material performance.

Palagi and colleagues' calculations suggest that the swimming speed would remain practically unchanged if the swimmers' size was reduced. If so, then a 5-μm-long rod (about a 200-fold reduction in size) would still swim at about 2–3 μm per second (comparable to half a body length per second) — slower than bacteria and paramecia, but on a par with the speeds achieved by other synthetic swimmers of similar size[1]. Advances in high-performance photoresponsive materials will be needed to further boost the microrobots' swimming speed. ■

Igor S. Aranson *is in the Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA.*
e-mail: aronson@anl.gov

1. Palagi, S. *et al. Nature Mater.* http://dx.doi.org/10.1038/nmat4569 (2016).
2. Paxton, W. F. *et al. J. Am. Chem. Soc.* **126,** 13424–13431 (2004).
3. Gao, W., Sattayasamitsathit, S., Orozco, J. & Wang, J. *J. Am. Chem. Soc.* **133,** 11862–11864 (2011).
4. Sanchez, S., Ananth, A. N., Fomin, V. M., Viehrig, M. & Schmidt, O. G. *J. Am. Chem. Soc.* **133,** 14860–14863 (2011).
5. Dreyfus, R. *et al. Nature* **437,** 862–865 (2005).
6. Ghosh, A. & Fischer, P. *Nano Lett.* **9,** 2243–2245 (2009).
7. Chang, S. T., Paunov, V. N., Petsev, D. N. & Velev, O. D. *Nature Mater.* **6,** 235–240 (2007).
8. Snezhko, A., Belkin, M., Aranson, I. S. & Kwok, W.-K. *Phys. Rev. Lett.* **102,** 118103 (2009).
9. Snezhko, A. & Aranson, I. S. *Nature Mater.* **10,** 698–703 (2011).
10. Cheang, U. K. & Kim, M. J. *J. Nanopart. Res.* **17,** 145 (2015).
11. Medina-Sánchez, M., Schwarz, L., Meyer, A. K., Hebenstreit, F. & Schmidt, O. G. *Nano Lett.* **16,** 555–561 (2016).
12. Katsu-Kimura, Y., Nakaya, F., Baba, S. A. & Mogami, Y. *J. Exp. Biol.* **212,** 1819–1824 (2009).
13. Camacho-Lopez, M., Finkelmann, H., Palffy-Muhoray, P. & Shelley, M. *Nature Mater.* **3,** 307–310 (2004).

VIROLOGY

The X-Files of hepatitis B

The HBx protein of hepatitis B virus has been found to co-opt a host-cell enzyme that targets the Smc5/6 protein complex for degradation. The finding identifies Smc5/6 as a cellular antiviral factor. SEE LETTER P.386

T. JAKE LIANG

Like special agents Mulder and Scully in the science-fiction television show *The X-Files*, the hepatitis B virus research community has for decades been chasing its own unsolved mystery — the truth about the enigmatic 'X' protein of this virus. The HBx protein was first shown to be involved in transcriptional activation, and has since been implicated in diverse cellular pathways, including signal transduction, apoptotic cell death, cell-cycle regulation and DNA repair. But how HBx exerts its effects has remained unclear. In this issue, Decorsière *et al.*[1] (page 386) provide intriguing evidence that HBx mediates the degradation of a host antiviral (restriction) factor by interacting with the ubiquitin–proteasome system — the major protein-degradation system in cells.

Hepatitis B virus (HBV) is a small DNA virus that has a partially double-stranded genome and replicates through an RNA intermediate. After entry into a host cell, the genome is converted to covalently closed circular DNA (cccDNA) that exists as a minichromosome in the nucleus and serves as the template for viral gene transcription. Viruses that are related to the HBV that infects humans have been discovered in many species, including ducks, woodchucks, squirrels and diverse species of bat. The virus predominantly infects liver cells, and can cause chronic infection even in the presence of an intact immune response.

Much of the uncertainty about the function of HBx has been attributed to the limitations of the experimental models in which it has been studied, because most are based on non-infectious systems. However, it is clear that HBx is required for effective HBV infection *in vivo*; woodchuck hepatitis virus that harbours defects in the gene encoding this protein is poorly infective[2,3].

The discovery of Smc5/6 as a viral restriction factor adds to a list of cellular-defence mechanisms against DNA-viral pathogens.

Many host factors are known to interact with HBx. Among them, damaged DNA binding protein 1 (DDB1) was first identified using a genetic approach[4]. This interaction was subsequently validated by structural and functional studies. But how this seemingly unrelated DNA-damage-response protein is involved in the function of HBx remained unclear. Like many typical breakthroughs in science, advances in an unrelated field provided the connection.

The ubiquitin–proteasome system has garnered much attention because of its central role in many cellular processes[5], and its components and biochemical pathways are well understood. Among them is an E3 ligase enzyme named CRL4. This member of the Cullin-RING ubiquitin ligase family uses DDB1 as an adaptor protein with which to target specific proteins for degradation. Several viral-gene products are known to target the CRL4–DDB1 complex[6], suggesting that the ubiquitin–proteasome system may be a common cellular pathway exploited by viruses to ensure productive infection. Decorsière *et al.* set out to identify proteins targeted for destruction by the complex formed between CRL4, DDB1 and HBx.

Using a clever protein-interaction approach, the authors identified the Smc5/6 protein complex, which is involved in several aspects of chromosome biology, as one such target (Fig. 1). HBx redirects the enzymatic function of CRL4 to target Smc5/6 for ubiquitination — a modification that marks the protein for degradation. They then used genetic and biochemical methods to show that the Smc5/6 complex indeed associates with the HBV genome (probably with the cccDNA, although this is not shown definitively) to inhibit viral transcription. A previous study[7] of Cullin-RING ligases did not identify the Smc5/6 complex as a target of CRL4, so it is unclear whether the complex is a natural substrate of CRL4. It is possible that HBx alters the substrate specificity of CRL4 such that the enzyme targets cellular proteins, in addition to the Smc5/6 complex, for degradation.

The discovery of Smc5/6 as a viral restriction factor adds to a growing list of intrinsic mechanisms in the cellular defence arsenal against DNA-viral pathogens[8,9]. It seems that the Smc5/6 complex binds to and suppresses only extrachromosomal (episomal), not chromosomally integrated, HBV DNA. This episome-specific function is reminiscent of another class of antiviral factor, the APOBEC family, which specifically targets

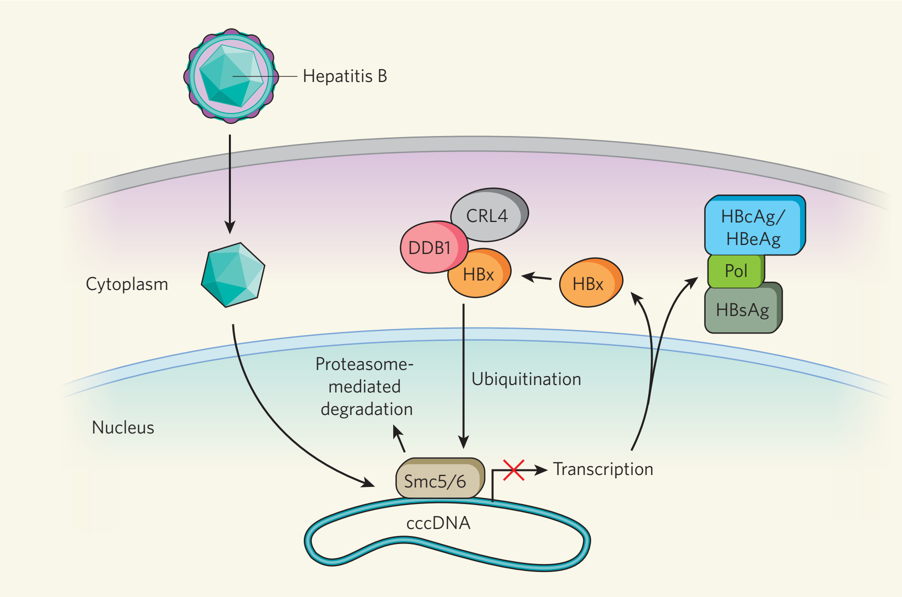


Figure 1 | Hepatitis B virus evasion of cellular antiviral function. After a hepatitis B virus (HBV) particle has entered a host cell (during which it becomes de-coated), its genome is converted to covalently closed circular DNA (cccDNA) that exists as a mini-chromosome in the nucleus and serves as the template for viral gene transcription. Three HBV proteins have well-defined functions: a core protein (named HBcAg, or HBeAg when secreted), a reverse transcriptase enzyme (Pol) and an envelope protein (HBsAg). Decorsière *et al.*[1] reveal that another viral protein, HBx, acts to degrade a cellular antiviral factor, the Smc5/6 protein complex. The authors show that Smc5/6 probably binds to the HBV cccDNA and thus inhibits viral transcription. But HBx interacts with DDB1, an adaptor protein for the cell's CRL4 E3 ubiquitin ligase enzyme complex, and this results in the Smc5/6 complex being targeted for ubiquitination — a modification that designates the complex for degradation by the cell's proteasome machinery.

episomal foreign DNA for modification and degradation[10]. It has been suggested that the APOBEC3A protein, whose expression is induced by interferon signalling proteins, binds to and edits HBV cccDNA, leading to cccDNA degradation[11]. Such pathways reveal that diverse cellular mechanisms have evolved to defend against HBV infection.

The Smc5/6 complex has been implicated in cell-cycle progression, chromosome organization and DNA repair[12], but little is known about its involvement in transcriptional regulation. Decorsière *et al.* have shown that this complex may have a key antiviral role by binding to viral genomes and silencing their transcription. How Smc5/6 targets episomal DNA to exert this silencing effect, and whether it has a similar activity against other DNA viruses, remain to be investigated. There is evidence[13] that HBV transcription is tightly regulated by epigenetic modifications (those that alter gene expression without modifying the nucleic-acid sequence), and it seems that HBx epigenetically modifies the HBV mini-chromosome. It is possible that the Smc5/6 complex silences transcription by affecting the epigenetic status of the viral mini-chromosome.

It has been suggested that HBx is present on the HBV mini-chromosome[10], although it is not known whether its association with the genome is necessary for the targeted degradation of the Smc5/6 complex. It is also unclear whether HBx interacts directly with Smc5/6; because the complex contains many proteins, it could be that a yet-unidentified factor in Smc5/6 is the direct target of the CRL4–DDB1–HBx complex.

Another remaining puzzle is why the gene that encodes HBx exists in the mammalian but not in avian hepatitis viruses. Did the interaction of this protein with Smc5/6 emerge as mammalian HBV diverged from its avian counterparts? It is intriguing to speculate that the HBx gene might have been acquired from the host as the virus entered mammals more than 10,000 years ago; acquisition of host genes is a typical evolutionary event for many viruses. It seems likely that Decorsière and colleagues' discovery is not the final episode of the 'HBx files'. ■

T. Jake Liang *is in the Liver Diseases Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-1800, USA.*
e-mail: jliang@nih.gov

1. Decorsière, A. *et al. Nature* **531,** 386–389 (2016).
2. Zoulim, F., Saputelli, J. & Seeger, C. *J. Virol.* **68,** 2026–2030 (1994).
3. Zhang, Z., Torii, N., Hu, Z., Jacob, J. & Liang, T. J. *J. Clin. Invest.* **108,** 1523–1531 (2001).
4. Lee, T.-H., Elledge, S. J. & Butel, J. S. *J. Virol.* **69,** 1107–1114 (1995).
5. Bosu, D. R. & Kipreos, E. T. *Cell Div.* **3,** 7 (2008).
6. Randow, F. & Lehner, P. J. *Nature Cell Biol.* **11,** 527–534 (2009).
7. Emanuele, M. J. *et al. Cell* **147,** 459–474 (2011).
8. Tavalai, N. & Stamminger, T. *Virus Res.* **157,** 128–133 (2011).
9. Wiebe, M. S. & Jamin, A. *J. Virol.* http://dx.doi.org/10.1128/JVI.00178-16 (2016).
10. Stenglein, M. D., Burns, M. B., Li, M., Lengyel, J. & Harris, R. S. *Nature Struct. Mol. Biol.* **17,** 222–229 (2010).
11. Lucifora, J. *et al. Science* **343,** 1221–1228 (2014).
12. Kegel, A. & Sjögren, C. *Cold Spring Harb. Symp. Quant. Biol.* **75,** 179–187 (2010).
13. Levrero, M. *et al. J. Hepatol.* **51,** 581–592 (2009).

CORAL REEFS

Turning back time

An *in situ* experiment finds that reducing the acidity of the seawater surrounding a natural coral reef significantly increases reef calcification, suggesting that ocean acidification may already be slowing coral growth. SEE LETTER P.362

JANICE M. LOUGH

In 1770, the only external threat to Australia's Great Barrier Reef was localized damage caused by European sailing ships such as HMS *Endeavour*, which struck the reef on 11 June that year. Wind the clock forward almost 250 years and the threats to tropical coral reefs worldwide have escalated to a level that imperils the survival of these complex, diverse and beautiful ecosystems[1]. On page 362 of this issue, Albright *et al.*[2] report an elegant field experiment in which they measured the response of a coral-reef community in the southern Great Barrier Reef to ocean chemistry conditions that were characteristic of the pre-industrial era. By turning back time in this way, they demonstrate that, all else being equal, net coral-reef calcification would have been around 7% higher than at present, suggesting that ocean acidification may already be diminishing coral-reef growth.

Threats to coral reefs range from regional factors, such as overfishing and land-based pollution, to global-scale issues arising from human interference in Earth's climate system, causing warming and acidification of the oceans. Tropical coral reefs have shown their sensitivity and vulnerability to the relatively modest amount of warming that has already occurred. This warming has, for example, resulted in mass coral-bleaching episodes

KEN CALDEIRA

on many of the world's reefs, most dramatically during the powerful 1997–98 El Niño event, and again during the equally large 2015–16 El Niño. Alongside this warming is the ongoing acidification of the oceans, caused by increased uptake of carbon dioxide from the atmosphere. Acidification is projected to increasingly compromise the ability of calcifying marine organisms to produce the calcium carbonate that forms their skeletons and shells[3].

The skeletons of coral animals form the backbone of tropical coral reefs and provide habitat for many thousands of other organisms. Calcification rates must be fast enough to withstand the natural forces of physical and biological erosion. Coral animals achieve this with the aid of the extra food and energy provided by single-celled symbiotic algae called zooxanthellae. These algae provide the beautiful colours of corals, but under stressful environmental conditions, such as unusually warm water temperatures, the algae are lost from the coral tissue, resulting in bleaching.

Corals build their skeletons by precipitating calcium and carbonate ions from seawater, but the absorption of CO_2 by seawater shifts the equilibrium away from carbonate-ion availability towards bicarbonate ions. The world's oceans absorb about 25% of the extra CO_2 that humans inject into the atmosphere — without this sink the planet would have warmed more than it has so far. But this absorption has decreased the pH, carbonate-ion concentration and saturation state of carbonate minerals of the surface oceans, collectively referred to as ocean acidification[4].

Since alarm bells starting ringing about the probable consequences of these changes for tropical coral reefs[5], numerous laboratory and field experiments have confirmed that calcification rates decline with ocean acidification. However, these studies have focused primarily on single coral species in experimental environments manipulated to mimic future conditions. The beauty and novelty of Albright and colleagues' study is that they restored the ocean chemistry of a natural reef to that of pre-industrial times, thus factoring out other potentially confounding factors, such as temperature.

The authors took advantage of the unusual hydrodynamic environment at One Tree Reef in the southern Great Barrier Reef (Fig. 1). Here, three lagoons are isolated from the ocean at low tide, but water flows between two of the lagoons (as a result of elevation differences) across a well-developed reef flat. For 60 minutes at low tide each day for 22 days, the researchers pumped a non-reactive dye solution across the reef flat, which for 15 days was enriched with sodium hydroxide to increase the alkalinity, and thus reduce the acidity, of the water.

Figure 1 | One Tree Reef in the southern Great Barrier Reef, Australia. The reef includes three lagoons in which water is separated from the rest of the ocean at low tide each day. This interval allowed Albright *et al.*[2] to study the effect of reducing the acidity of the water on coral growth.

This application of dual tracers — alkalinity as an active tracer and the dye as a passive tracer — has been previously applied to atmospheric aerosols. The tracing allowed the authors to calculate how much of the added alkalinity was taken up by the reef-flat community (around 17%). The authors also recorded an increase in aragonite saturation state (a measure of the availability of dissolved carbonate and calcium ions) of around 0.6 and an approximately 7% increase in net coral-community calcification, findings that closely match results from previous experimental manipulations of ocean chemistry[6].

One implication of these discoveries is that ocean acidification may already be contributing to observed declines in coral growth[7]. As Albright *et al.* acknowledge, ocean chemistry is not the only influence on coral calcification. The surface temperatures of the tropical oceans were also significantly cooler in pre-industrial times[8]. Although current rates of warming are deemed detrimental, some corals may have initially benefited from warmer oceans[7]. We need to consider not just what ocean acidification and warming separately mean for coral-reef calcification, but also their combined effects and how these factors interact with local stressors, as well as the consequences for coral-reef communities as a whole. Extension of Albright and colleagues' approach to include other factors could help us to get to grips with the relative risks of a rapidly changing future for different coral reefs and locations.

By turning back time in an experimental scenario, Albright *et al.* have demonstrated the vulnerability of the process of reef calcification to ocean acidification. But we cannot turn back time for the world's tropical coral-reef ecosystems; we have already committed them to a warmer and more acidic future that is likely to fundamentally change them from the diverse and beautiful ecosystems of the pre-industrial era to much simpler ecosystems that are not dominated by corals[9]. Such a loss will also affect many millions of people who depend on coral reefs for their livelihoods and food security.

The 2015 Paris climate agreement to keep the average global temperature "well below 2 °C above pre-industrial levels" contends that this limit "would significantly reduce the risks and impacts of climate change". But a 2 °C limit may not be enough for tropical coral reefs, which are at high risk of substantial impacts from ocean warming and acidification, even with the most stringent emission targets[10]. ■

Janice M. Lough *is at the Australian Institute of Marine Science and the Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland, Australia.*
e-mail: j.lough@aims.gov.au

1. Hoegh-Guldberg, O. *et al. Science* **318,** 1737–1742 (2007).
2. Albright, R. *et al. Nature* **531,** 362–365 (2016).
3. Doney, S. C., Fabry, V. J., Feely, R. A. & Kleypas, J. A. *Annu. Rev. Mar. Sci.* **1,** 169–192 (2009).
4. Jiang, L.-Q. *et al. Glob. Biogeochem. Cycles* **29,** 1656–1673 (2015).
5. Kleypas, J. A. *et al. Science* **284,** 118–120 (1999).
6. Chan, N. C. & Connolly, S. R. *Glob. Change Biol.* **19,** 282–290 (2013).
7. Lough, J. M. & Cantin, N. E. *Biol. Bull.* **226,** 187–202 (2014).
8. Tierney, J. E. *et al. Paleoceanography* **30,** 226–252 (2015).
9. Fabricius, K. E. *et al. Nature Clim. Change* **1,** 165–169 (2011).
10. Gattuso, J.-P. *et al. Science* **349,** 45 (2015)

This article was published online on 24 February 2016.

doi:10.1038/nature17148

Interface dynamics and crystal phase switching in GaAs nanowires

Daniel Jacobsson[1,2], Federico Panciera[3,4], Jerry Tersoff[4], Mark C. Reuter[4], Sebastian Lehmann[1], Stephan Hofmann[3], Kimberly A. Dick[1,2] & Frances M. Ross[4]

Controlled formation of non-equilibrium crystal structures is one of the most important challenges in crystal growth. Catalytically grown nanowires are ideal systems for studying the fundamental physics of phase selection, and could lead to new electronic applications based on the engineering of crystal phases. Here we image gallium arsenide (GaAs) nanowires during growth as they switch between phases as a result of varying growth conditions. We find clear differences between the growth dynamics of the phases, including differences in interface morphology, step flow and catalyst geometry. We explain these differences, and the phase selection, using a model that relates the catalyst volume, the contact angle at the trijunction (the point at which solid, liquid and vapour meet) and the nucleation site of each new layer of GaAs. This model allows us to predict the conditions under which each phase should be observed, and use these predictions to design GaAs heterostructures. These results could apply to phase selection in other nanowire systems.

Many materials can grow in multiple (meta)stable crystal structures, and phase selection is one of the most fundamental problems in materials science. However, the selection process is difficult to access experimentally; for example, many metastable phases are obtained only by rapid quenching of a liquid into a polycrystalline multiphase solid. By contrast, nanowires provide an ideal system for studying phase selection. Typical zinc-blende (ZB)-structure III-V semiconductors form nanowires in the wurtzite (WZ) structure as well as the ZB[1–5]. Nanowires can easily be switched between these phases by varying the temperature, source-material flux or impurities[3,4,6–14]; and their small diameter guarantees that they are single crystals, with phase switching occurring along the growth axis where it is easily observed. ZB and WZ semiconductors have different band structures[15], which creates opportunities for designing modulated nanowire structures with new electronic properties. Crystal phase heterostructures are particularly interesting because they can access the electronic properties of heterostructure quantum dots for photonics and single-electron-transistor applications, but without the challenge of achieving compositional control[16–19].

To take full advantage of the possibilities offered by crystal structure control, a detailed understanding of the physics behind crystal phase selection is required. On the basis of post-growth observations, different models have been proposed for phase selection. These models emphasize the role of supersaturation, catalyst geometry and interfacial energies[20–25]. Experimental results are typically interpreted in terms of the dominant role of one of these factors[6–8,10,12,20,26–28].

Here, we directly observe the dynamic processes that take place during nanowire growth for each crystal phase, and during the switch between phases, using *in situ* transmission electron microscopy. We find surprising differences in the structure and dynamics during growth of ZB and WZ nanowires. The switching process itself, and the associated changes in geometry, provide clues that allow us to develop a new model identifying the underlying mechanism driving crystal phase selection. In this model, droplet geometry is the key parameter in determining structure, but in an indirect way, via its effect on the nanowire edge morphology. This understanding allows us to form crystal phase quantum dots with atomic layer precision.

Imaging interface and catalyst geometry

We observed the two GaAs nanowire crystal phases during growth *in situ* using a Hitachi H-9000 ultra-high vacuum transmission electron microscope (UHVTEM)[29,30]. Si substrates were first covered with pre-grown GaAs nanowires using standard metal–organic vapour phase epitaxy (MOVPE) and Au aerosol particles with diameters of 30 nm, 50 nm and 70 nm. Such samples were loaded into the TEM and heated resistively, using a pyrometer to calibrate the temperature at each heating current (see Methods). Pure trimethylgallium (TMGa) and arsine (AsH_3) were used as precursor gases, and were introduced close to the substrate using separate capillary tubes to a maximum total pressure during imaging of 2×10^{-5} Torr. Details of how the growth parameters *in situ* compare to conventional MOVPE are provided in Methods. On heating to temperatures of about 550 °C, a liquid AuGa droplet formed at the nanowire tips and growth took place at the droplet/nanowire interface. Growth was recorded at 30 images per second. Dark-field imaging conditions, as used in Fig. 1, allow the crystal structure to be distinguished, that is, the WZ phase and the two twin variants of the ZB phase (Extended Data Fig. 1). Bright-field imaging conditions, as used in Fig. 2, allow a more accurate determination of the dimensions of the droplet and the nanowire.

We find that, *in situ*, both ZB and WZ GaAs can be grown by varying the precursor pressures (the V/III ratio) while maintaining a constant temperature. Within the parameter range accessible *in situ*, WZ GaAs forms at higher V/III ratios and ZB at lower ratios at steady-state conditions; transient conditions are discussed below. The two phases show marked differences in terms of their growth dynamics. For WZ GaAs, growth proceeds by step flow across the droplet/nanowire interface (Fig. 1a and Supplementary Video 1). Steps flow slowly with each one starting as soon as the previous one has completed its flow (Extended Data Fig. 2a). By counting the number of step-flow events and correlating with the length of the nanowire (Extended Data Fig. 2b), as in ref. 31, we find that each step flow represents the addition of one WZ GaAs(0001) bilayer, with a height of 0.3 nm. The growth rates are low under the conditions accessible *in situ*, typically one bilayer per minute (see Methods and Supplementary Video 1). Growth rates are proportional to AsH_3 pressure (Extended Data Fig. 2c), which suggests

[1]Solid State Physics and NanoLund, Lund University, Box 118, 221 00 Lund, Sweden. [2]Centre for Analysis and Synthesis, Lund University, Box 124, 221 00 Lund, Sweden. [3]Department of Engineering, University of Cambridge, 9 JJ Thomson Avenue, Cambridge CB3 0FA, UK. [4]IBM T. J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, New York 10598, USA.

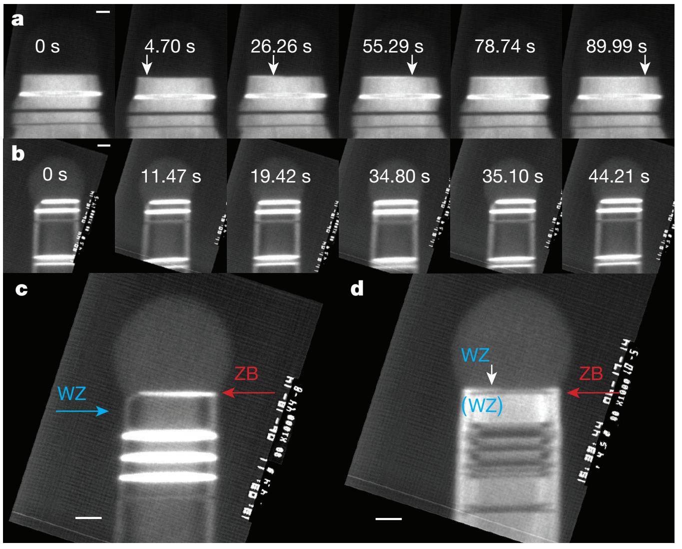


Figure 1 | Interface dynamics during WZ and ZB growth of GaAs. **a**, Images extracted from a dark-field movie recorded during WZ GaAs growth. A step flows across the top facet of a nanowire that has a diameter of 60 nm; the position of the step is indicated by the arrows. See also Supplementary Video 1. Growth conditions: 550 °C, AsH_3 pressure of 1×10^{-5} Torr, TMGa pressure of 3.5×10^{-8} Torr. The narrow stripes show previously grown ZB segments: ZB can occur in two twinned orientations, one appearing bright and the other dark in this imaging condition. **b**, Images extracted from a dark-field movie recorded during ZB GaAs growth. The truncation slowly fills then jumps back to its maximum size (the simultaneous rapid step flow across the growth interface is not visible). See also Supplementary Video 2. Growth conditions (not steady-state for this example): 550 °C, AsH_3 pressure increased from 10^{-7} Torr to 1.4×10^{-5} Torr a few seconds before the first image was recorded, TMGa pressure of 2.0×10^{-8} Torr. ZB phases appear bright and WZ dark in this imaging condition. Note that the truncation shows strongly on one side of the nanowire; this was typical (see text), although some (5%) nanowires showed synchronized oscillation on both sides, as in Supplementary Video 5. The relative time of each image in **a** and **b** is shown in seconds. **c**, The first ZB bilayer growing on WZ, with imaging conditions as in **b**. The AsH_3 pressure was reduced; the droplet is in the process of growing larger past the critical volume; the first layer of ZB appears followed immediately by a truncation that cuts into the previously grown WZ. **d**, The first WZ bilayer growing on a ZB segment. The AsH_3 pressure was increased; the droplet is in the process of growing smaller; the truncation fills in and the first layer of WZ appears via step flow. The truncation does not appear, and slow step flow occurs, even though ZB covers the top facet. Scale bars are 10 nm in all images.

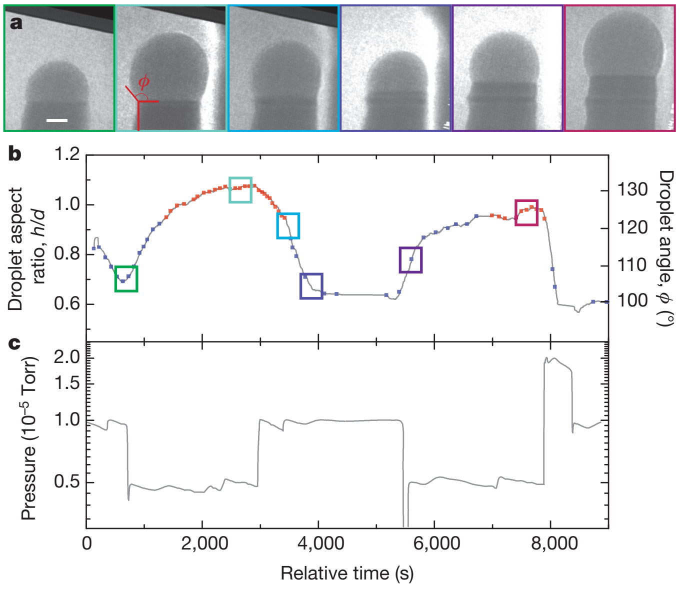


Figure 2 | Changes in droplet volume during phase switching. **a**, Series of bright-field images obtained during growth of a GaAs nanowire at varying AsH_3 pressure, constant TMGa pressure (2×10^{-8} Torr) and constant temperature (550 °C). Scale bar is 10nm. **b**, Droplet aspect ratio h/d and angle ϕ as defined in **a**. The times of the images in **a** are indicated by coloured boxes. The droplet volume takes several minutes to respond to the change in pressure. The crystal phase is indicated by blue (WZ) and red (ZB) squares. Each red square marks the occurrence of a truncation of the top facet and nucleation of a ZB bilayer. Each blue square marks the identification of WZ growth via step flow. **c**, AsH_3 pressure variation over time.

that growth under these circumstances is limited by the arrival and incorporation of As (see Methods). We can then understand the step-flow dynamics through the solubility of As in AuGa, which is generally accepted to be low[32]. Since the droplet contains no reservoir of the rate-limiting species (in our case, As), the arriving atoms are incorporated immediately into the nanowire, leading to slow and gradual step flow[33].

The growth of ZB GaAs looks quite different (Fig. 1b, Supplementary Video 2). Growth similarly proceeds by addition of bilayers, but each bilayer flows across the growth interface too rapidly to observe. Furthermore, the droplet/nanowire interface shows an oscillating geometry at the trijunction (at which solid, liquid and vapour meet) that is similar to that seen in Si, ZB GaP, Ge and Al_2O_3 (refs 30, 31, 34, 35). The edge of the nanowire appears truncated (Fig. 1b, first panel). The three-dimensional geometry of this 'edge facet' is shown schematically in Extended Data Fig. 1. Material gradually adds to the edge facet to fill in the corner (Fig. 1b, second and third panels). The interface jumps forwards as one step flows quickly, and the edge facet reappears (Fig. 1b, fourth and fifth panels).

Each oscillation in trijunction geometry is correlated with the nucleation and flow of a new bilayer, as in ref. 31. The step moves quickly, even with no reservoir of As, because it is supplied by material from the truncated volume. A rough estimate of the change in the truncated volume is consistent with it being the source of the one bilayer of growth.

In parallel with the observed changes in the interface dynamics, the droplet geometry also changes as we vary the V/III ratio to achieve growth of WZ and ZB GaAs. Figure 2 and Supplementary Video 3 show the effect on the droplet of changing the AsH_3 pressure between high values (to achieve WZ) and low values (to achieve ZB) at constant TMGa pressure (2×10^{-8} Torr). Temperature was kept constant throughout the experiments (at 540–560 °C) to avoid introducing temperature dependencies that would obscure the observed trends (see Methods). Throughout the changes in AsH_3 pressure, the nanowire continued to grow. The most obvious feature is the change in the volume of the droplet. We quantify this in Fig. 2b via the droplet aspect ratio h/d (droplet height divided by nanowire diameter at the growth interface) or, equivalently, via the droplet angle ϕ (angle between the basal plane and the tangent to the droplet at its edge; see Fig. 2a). On decreasing the AsH_3 pressure, the droplet increases in volume; increasing the AsH_3 pressure decreases the droplet volume. A quasi-steady-state volume is reached, depending on the V/III ratio. These volume changes must be driven by the addition or subtraction of Ga: we do not expect Au to move in and out of the droplet, because its diffusion on GaAs is assumed to be negligible at this temperature[36], whereas As makes up only a small fraction of the volume, owing to its low solubility (discussed above). An Au–Ga–As alloy with over 40% Ga forms a liquid at this temperature, with no upper limit on the Ga content[32]; consequently, droplets of a range of volumes are possible. Furthermore, because the volume changes occur more quickly than the rate at which Ga could be consumed by incorporation into the growing nanowire, the Ga must be supplied or removed

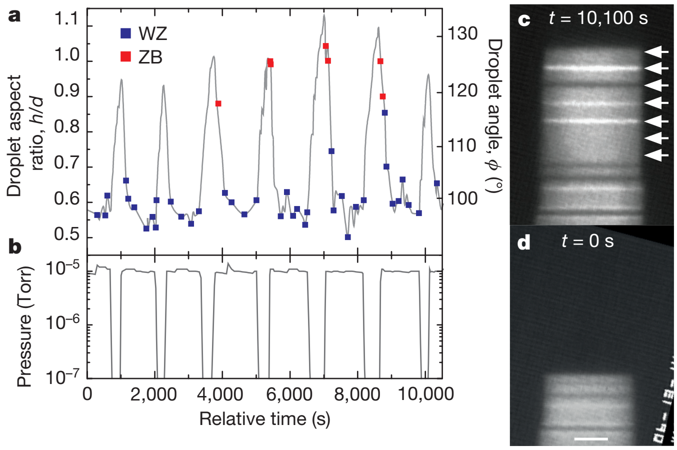


Figure 3 | Growth of a WZ nanowire containing multiple narrow ZB segments. TMGa pressure (2×10^{-8} Torr) and temperature (550 °C) were constant. **a**, Droplet aspect ratio h/d and angle ϕ versus time (grey line), and crystal phase versus time (squares). Each square indicates the addition of a bilayer of WZ (blue) or ZB (red), as described in Fig. 1. ZB segments with thicknesses of zero, one or two bilayers form each time h/d increases. **b**, AsH_3 pressure variation over time. The pressure was held at 1×10^{-5} Torr to grow WZ, but was pulsed downwards to less than 10^{-8} Torr for seven intervals with durations 5 min, 5 min, 7 min, 7 min, 9 min, 9 min and 5 min, respectively. **c**, **d**, Images of the nanowire at the start (**d**) and end (**c**) of the experiment; scale bar is 10 nm. In **c**, the position of the growth front at each of the seven intervals is indicated by an arrow. The ZB segments grown in the 7- and 9-min intervals are visible as narrow stripes. Three segments have one ZB twin orientation (bright contrast in this imaging condition) and one has the other ZB twin orientation (dark contrast).

by surface diffusion along the nanowire. In the simplest picture[30], there is a surface reservoir of mobile Ga adatoms that equilibrate with the droplet over time whenever the chemical potentials of Ga in the droplet (and on the surface) are changed by altering the AsH_3 pressure. In this way, the V/III ratio controls the droplet size. (A more complete treatment would include diffusion and the effect of As flux on Ga diffusion, but this would not change the general picture[22].)

We have shown above that crystal structure and droplet volume both change as the V/III ratio varies. We now explore how they correlate with each other. During a growth experiment, it is possible to measure crystal structure and droplet volume by alternating between dark- and bright-field imaging conditions. The crystal structure identified during the experiment in Fig. 2a is shown as the red and blue data points in Fig. 2b. It is clear that the switch between WZ and ZB crystal structure occurs as the droplet passes a certain aspect ratio—under these conditions, this is $h/d \approx 0.95$ and $\phi \approx 125°$. (Other experiments show a small hysteresis that is not visible in this data; see, for example, Supplementary Fig. 3.) Because the droplet takes several minutes to respond to the pressure change, it is clear that WZ–ZB growth is correlated to h/d and ϕ rather than to the instantaneous AsH_3 pressure. This direct correlation between crystal switch and droplet dimensions (volume, aspect ratio and angle), governed ultimately by the V/III ratio, is a key result that provides the basis for the model we develop below.

Understanding that the droplet geometry is the critical parameter, rather than the gas environment itself, provides useful guidance in growing crystal phase heterostructures. The length of each crystal phase segment depends on the time during which the droplet has the appropriate geometry. The relatively slow kinetics of the change in droplet volume mean that the V/III ratio must be designed with appropriate offsets in timing. An example is shown in Fig. 3. Here, the V/III ratio was set initially at a value that formed WZ GaAs. The AsH_3 pressure was then decreased, for short pulses, to a V/III ratio that would be expected to form ZB GaAs. The result is a series of ZB inclusions in a WZ nanowire with lengths that are repeatable, but not directly proportional to the pulse duration. The shortest pulses did not form ZB GaAs at all. Longer pulses produced one or two bilayers of ZB stacking. The correlation between droplet volume and crystal phase in Fig. 3 confirms that the droplet must reach a critical volume for the structural change to occur; the reduction of AsH_3 pressure in itself may not trigger a structure change. Designing a crystal phase heterostructure thus requires consideration of the kinetics of the change in droplet volume.

A model for interface geometry

Developing a framework to understand the relationship observed above between droplet volume, interface growth dynamics and crystal structure requires two additional key observations. The first observation is that when the droplet changes volume, it changes composition too. This could in principle affect phase selection, for example, by changing the surface energy[27]. It is therefore not immediately clear which factor determines crystal structure, the Au:Ga ratio or the geometry (h/d and ϕ). To establish this we measured h/d and ϕ at the switch for one particular nanowire, then allowed the nanowire to increase its diameter by conformal growth on the sidewalls, and again measured h/d and ϕ while inducing a switch (Extended Data Fig. 3). Because the nanowire widens, but the amount of Au present does not change, relatively more Ga is needed to achieve the same h/d and ϕ. We did not see any strong effect of diameter on the switch between crystal phases; h/d and ϕ appear to be the controlling parameters.

The second observation concerns the relationship between crystal phase and the dynamics at the growth front. In Fig. 1c, d and Supplementary Videos 2 and 4 we show the crystal switch in more detail, specifically the growth of the first ZB layer on WZ GaAs and the first WZ layer on ZB GaAs. Starting from a WZ nanowire, we reduce the AsH_3 pressure (Fig. 3c); the droplet enlarges and the first ZB layer forms as the critical h/d and ϕ are reached. As the ZB step flows (too rapidly to see) across the growth interface, an edge facet appears (Supplementary Video 4). This facet cuts into the WZ beneath. Thus, even though steady-state growth of WZ GaAs proceeds without an edge facet, it is possible to form an edge facet in WZ GaAs under appropriate conditions. Conversely, if we start from a ZB nanowire and increase the AsH_3 pressure, as shown in Fig. 1d, then the droplet shrinks and the first WZ layer grows. It grows by slow step flow and without an edge facet appearing (Supplementary Video 2), even though ZB is exposed on the growth interface as it starts to grow.

These morphology observations are particularly surprising. One might expect the presence of an edge facet to be controlled by the crystal structure at that facet; but, instead, Fig. 1c, d shows that it correlates with the crystal structure on the main (that is, ZB(111) or WZ(0001)) growth facet, and thus with h/d and ϕ. To determine cause and effect, we analyse the ways in which the droplet angle ϕ affects the morphology of the growth interface. Equilibrium crystal shapes generally do not have edge angles as sharp as 90°, so one would expect an edge facet in a GaAs crystal. A sharp edge only exists during nanowire growth because the droplet is present, providing a capillary force that can pull on the edge facet and shrink its size to zero. We calculate the circumstances under which this occurs. We assume an ideal, symmetric nanowire for which the droplet angle ϕ is the same all around the edge (see Fig. 4a); the more realistic, asymmetric case is discussed below. The difference in free energy between this ideal nanowire with an edge facet and a nanowire with the same geometry but a sharp edge is[31]

$$\Delta E = c_1 y L + \frac{1}{2} y^2 (\cot\theta)(\mu_{\rm cat} - \mu_0) L + c_2 y^2 L \qquad (1)$$

in which L is the total length of the edge, y is the facet length and θ is the facet angle (see Fig. 4a), $\mu_{\rm cat} - \mu_0$ reflects the supersaturation

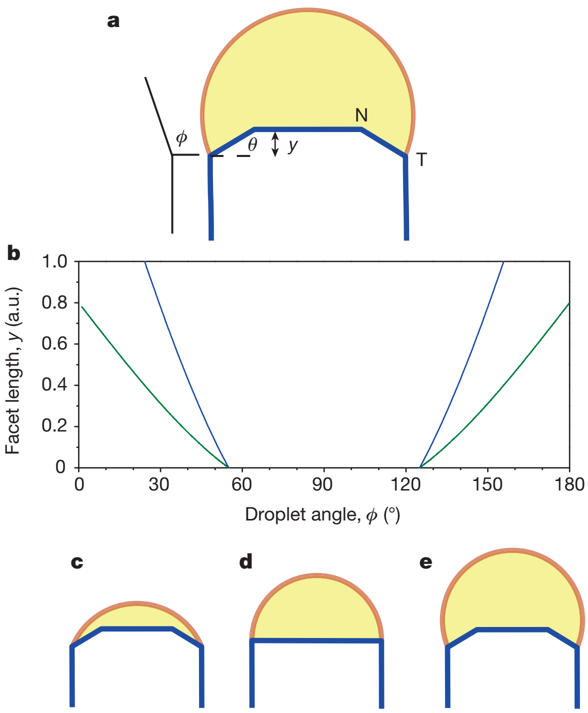


Figure 4 | Model relating droplet size to interface morphology. **a**, Schematic of a quasi-two-dimensional, ideal, symmetric nanowire illustrating the droplet angle ϕ, and the edge facet angle θ and length (in the growth direction) y. 'N' marks the interior point at which ZB nucleates; 'T' marks the trijunction. **b**, Edge facet length y versus droplet angle ϕ, calculated for two values of supersaturation (blue, low; green, high) for the symmetric nanowire. There is a range of ϕ in which $y = 0$; in this range $c_1 > 0$ and the edge facet does not exist. Supersaturation does not affect this range. **c**–**e**, Schematics of the nanowire and droplet for $\phi < 90°$ (**c**), $\phi = 90°$ (**d**) and $\phi > 90°$ (**e**). The possibility of the droplet depinning from the point T is not included in the model.

of chemical potential μ in the catalyst, and c_2 includes various other second-order terms[31]. For a sufficiently small facet length y, this energy is dominated by the linear term c_1, which reflects the capillary forces acting on the corner facet. Therefore, for $c_1 < 0$, it is always energetically favourable to have the edge facet ($\Delta E < 0$), whereas for $c_1 > 0$, we expect the edge to be sharp everywhere and have no facet. By examining c_1 in more detail and including all the capillary terms[37], we find

$$c_1 = \gamma_e \frac{1}{\sin\theta} - \gamma_{vs} - \gamma_{ls} \frac{\cos\theta}{\sin\theta} + \gamma_{vl} \sin\phi \quad (2)$$

in which γ_e is the liquid–solid interfacial energy at the edge facet, γ_{vs} is the vapour–solid interfacial energy on the sidewall, γ_{ls} is the liquid–solid interfacial energy at the main growth facet and γ_{vl} is the vapour–liquid interfacial energy.

The key point here is the presence of ϕ in equation (2). This implies that the droplet angle alters c_1 and, hence, changes the lowest-energy state of the nanowire from one with an edge facet to one with a sharp corner. A hemispherical droplet ($\phi = 90°$) yields the maximum possible value of c_1, and so is the most favourable for eliminating the edge facet. This is shown in Fig. 4b, in which we calculate the length y of the edge facet as a function of angle ϕ for a symmetric, but otherwise arbitrary, illustrative case. In our experiments, the droplet is never less than a hemisphere (ϕ is always greater than 90°) during stable growth. Thus, the analysis in equations (1) and (2) predicts that a switch could be observed between a large droplet with an edge facet (Fig. 4e) and a smaller droplet with a sharp edge (Fig. 4d); that is indeed what we observe.

Connecting geometry to crystal phase

We now consider the ways in which the presence or absence of an edge facet controls the crystal phase. Without attempting to develop a microscopic model, we can understand heuristically how this would occur by considering a previous analysis[20]. Several models have argued that the crystal structure should be determined by the location of the nucleation event on the main growth facet[6–8,10,13,20–23,26,38]. The argument is that the metastable WZ phase can only grow if it has a lower nucleation barrier than does the ZB phase. With sharp edges, nucleation is expected to occur at the trijunction (rather than in the middle of the facet), so the solid–vapour interface plays a critical part. In particular, the solid–vapour interface energy is thought to be lower for WZ nanowires, reducing the nucleation barrier for WZ relative to ZB in this geometry[20]. However, when edge facets are present, nucleation on the main facet occurs away from the trijunction (presumably at position 'N' in Fig. 4a; ref. 31) and the liquid–vapour interface plays no part. If we adopt this argument, then it is no surprise that the change in the trijunction geometry can result in easier nucleation of WZ than ZB in one case, but not the other. This argument is also qualitatively consistent with *in situ* X-ray diffraction studies that infer (indirectly) that crystal structure in GaAs nanowires is determined by the geometry of the liquid–vapour interface[39,40]. Here, we have not considered effects of interlayer interactions on the nucleation barrier, which might lead to formation of higher-order crystal phases, such as 4H, under certain conditions[41], because we do not observe such phases in our experiments.

At very small h/d, our analysis also suggests the possibility of an edge facet and, hence, ZB growth (Fig. 4b). Although we cannot access such conditions in our experiments, they could occur transiently at the beginning of nanowire growth, because the droplet has much a smaller h/d ratio when sitting on a flat surface[42], and perhaps at the end of growth if material in the droplet is consumed. Indeed, the ZB phase has been observed at the bases and tips of WZ nanowires[20], although under growth conditions that are different enough that our model might not be applicable. Recent experiments have demonstrated two transitions (from ZB to WZ and then back to ZB) as the V/III ratio is increased[9], consistent with the model in Fig. 4. However, because our experiments have only a limited range of V/III ratios, we cannot observe the second transition back to ZB and so cannot assess whether the transition is associated with interface structure in a way that is analogous to the switch at lower group V pressures.

The discussion above is simplified in several respects. No difference in interfacial or surface energies between ZB and WZ structures is included. The small, but real, differences could lead to hysteresis in the switching angle as the droplet grows and shrinks. However, data such as that in Fig. 2, which displays no strong hysteresis, suggest that the droplet angle has a larger effect than do the differences between ZB and WZ interfacial energies.

More importantly, our quasi-two-dimensional model treats the droplet angle ϕ as uniform all the way around the edge. For the true three-dimensional geometry, in which the droplet sits on a hexagonal prism whose side lengths may not be equal[43], it is clear that ϕ will vary around the trijunction. Suppose that for a WZ nanowire we change the conditions to enlarge the droplet. At some point, ϕ will become large enough along one edge for that edge to become truncated, even though other edges remain sharp. As the droplet continues to grow, every other edge will progressively become truncated. Therefore, we expect any nanowire with unequal edge lengths to exhibit a mix of sharp and truncated edges over a range of droplet volumes. In the experiments, we observe the ZB phase once the first truncated corner appears (Fig. 1 and Supplementary Videos 2 and 4). Because we typically stabilize the conditions as soon as we see the crystal switch, the majority of nanowires presented here show a mix of sharp and truncated edges. However, we also observe symmetrically oscillating nanowires (Supplementary Video 5), presumably because the wire is more symmetric or because the droplet has grown

large enough to cause all edges to be truncated. The observation mentioned above that the ZB phase grows if any edge is truncated, whereas WZ grows only when all edges are sharp, implies that nucleation of ZB at a truncated edge is actually easier than nucleation of either WZ or ZB at a sharp one. This somewhat unexpected result could provide guidance in refining model parameters in nucleation calculations.

Conclusions

Direct observation during growth has enabled us to probe the phenomena controlling crystal phase in nanowires. WZ and ZB crystal phases in GaAs appear markedly different during growth in terms of the morphology of the nanowire/droplet interface, the flow of steps and the droplet size. The step-flow kinetics can be understood as a consequence of As-limited growth, low As solubility in the droplet, and the role of the edge truncation as an alternative reservoir. Examining the switch between phases suggests a scenario in which the growth conditions (here, the V/III ratio) determine the volume of the droplet and, hence, its aspect ratio h/d and angle ϕ; the value of ϕ determines whether an edge facet will be present; the presence or absence of the edge facet determines the nucleation site for a new layer; and the nucleation site determines which phase, WZ or ZB, is most likely to nucleate. Our interpretation differs markedly from previous models of phase selection. Because nanowire growth has been achieved using a wide range of parameters and growth techniques, it is possible that phase selection is controlled by different physics under different circumstances. However, the regime we analyse here, MOVPE under As-limited conditions, has advantages for atomic-level control and high-throughput manufacturing.

This understanding of the causal sequence, in particular the changes in droplet volume with conditions and the controlling role of the droplet angle ϕ, has practical consequences. First, the large changes in droplet volume as a function of conditions may be relevant to aspects of nanowire growth other than crystal phase control[13]. For example, kinking can be caused by depinning of droplets from the nanowire tip[37,44], and experiments such as those shown here can explore the range of conditions under which droplets attain sizes that are sufficiently large or sufficiently small to cause depinning. In terms of crystal phase control, because the Au:Ga ratio in the droplet seems not to be critical, our results might be applicable to self-catalysed (Au-free) nanowire growth (although any small difference in liquid surface energies could lead to a slightly different critical angle). A second consequence is that any means of controlling the energy balance between a truncated and sharp edge should affect the crystal phase: we used the V/III ratio here, but temperature and surfactants are other possible ways to tune the crystal phase. We anticipate that similar behaviour may occur in other III-V semiconductors that exhibit polytypism, although it is not guaranteed because the various interfacial-energy parameters are material-specific. Finally, understanding how crystal structure switching depends on the kinetics of group III motion into and out of the droplet helps us work towards precise control of individual crystal phase superlattices, to enable fabrication of new types of electronic devices that make full use of the possibilities for engineering band structure that are provided by crystal-phase-engineered nanowires.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 October 2015; accepted 6 January 2016.

1. Krogstrup, P. *et al.* Structural phase control in self-catalyzed growth of GaAs nanowires on silicon (111). *Nano Lett.* **10,** 4475–4482 (2010).
2. Ikejiri, K., Kitauchi, Y., Tomioka, K., Motohisa, J. & Fukui, T. Zinc blende and wurtzite crystal phase mixing and transition in indium phosphide nanowires. *Nano Lett.* **11,** 4314–4318 (2011).
3. Husanu, E., Ercolani, D., Gemmi, M. & Sorba, L. Growth of defect-free GaP nanowires. *Nanotechnology* **25,** 205601 (2014).
4. Caroff, P. *et al.* Controlled polytypic and twin-plane superlattices in III–V nanowires. *Nature Nanotechnol.* **4,** 50–55 (2009).
5. Koguchi, M., Kakibayashi, H., Yazawa, M., Hiruma, K. & Katsuyama, T. Crystal structure change of GaAs and InAs whiskers from zinc-blende to wurtzite type. *Jpn J. Appl. Phys.* **31,** 2061–2065 (1992).
6. Wallentin, J. *et al.* Changes in contact angle of seed particle correlated with increased zincblende formation in doped InP nanowires. *Nano Lett.* **10,** 4807–4812 (2010).
7. Lehmann, S., Wallentin, J., Jacobsson, D., Deppert, K. & Dick, K. A. A general approach for sharp crystal phase switching in InAs, GaAs, InP, and GaP nanowires using only group V flow. *Nano Lett.* **13,** 4099–4105 (2013).
8. Joyce, H. J., Wong-Leung, J., Gao, Q., Tan, H. H. & Jagadish, C. Phase perfection in zinc blende and wurtzite III–V nanowires using basic growth parameters. *Nano Lett.* **10,** 908–915 (2010).
9. Lehmann, S., Jacobsson, D. & Dick, K. A. Crystal phase control in GaAs nanowires: opposing trends in the Ga- and As-limited growth regimes. *Nanotechnology* **26,** 301001 (2015).
10. Algra, R. E. *et al.* The role of surface energies and chemical potential during nanowire growth. *Nano Lett.* **11,** 1259–1264 (2011).
11. Dheeraj, D. L. *et al.* Controlling crystal phases in GaAs nanowires grown by Au-assisted molecular beam epitaxy. *Nanotechnology* **24,** 015601 (2013).
12. Johansson, J. *et al.* Effects of growth conditions on the crystal structure of gold-seeded GaP nanowires. *J. Cryst. Growth* **310,** 5102–5105 (2008).
13. Yuan, X. *et al.* Tunable polarity in a III–V nanowire by droplet wetting and surface energy engineering. *Adv. Mater.* **27,** 6096–6103 (2015).
14. Xu, T. *et al.* Faceting, composition and crystal phase evolution in III–V antimonide nanowire heterostructures revealed by combining microscopy techniques. *Nanotechnology* **23,** 095702 (2012).
15. De, A. & Pryor, C. E. Predicted band structures of III–V semiconductors in the wurtzite phase. *Phys. Rev. B* **81,** 155210 (2010).
16. Akopian, N., Patriarche, G., Liu, L., Harmand, J. C. & Zwiller, V. Crystal phase quantum dots. *Nano Lett.* **10,** 1198–1201 (2010).
17. Vainorius, N. *et al.* Confinement in thickness-controlled GaAs polytype nanodots. *Nano Lett.* **15,** 2652–2656 (2015).
18. Assali, S. *et al.* Direct band gap wurtzite gallium phosphide nanowires. *Nano Lett.* **13,** 1559–1563 (2013).
19. Dick, K. A., Thelander, C., Samuelson, L. & Caroff, P. Crystal phase engineering in single InAs nanowires. *Nano Lett.* **10,** 3494–3499 (2010).
20. Glas, F., Harmand, J. C. & Patriarche, G. Why does wurtzite form in nanowires of III–V zinc blende semiconductors? *Phys. Rev. Lett.* **99,** 146101 (2007).
21. Dubrovskii, V. G., Sibirev, N. V., Harmand, J. C. & Glas, F. Growth kinetics and crystal structure of semiconductor nanowires. *Phys. Rev. B* **78,** 235301 (2008).
22. Dubrovskii, V. G. Influence of the group V element on the chemical potential and crystal structure of Au-catalyzed III–V nanowires. *Appl. Phys. Lett.* **104,** 053110 (2014).
23. Krogstrup, P. *et al.* Advances in the theory of III–V nanowire growth dynamics. *J. Phys. D* **46,** 313001 (2013).
24. Johansson, J. *et al.* Effects of supersaturation on the crystal structure of gold seeded III−V nanowires. *Cryst. Growth Des.* **9,** 766–773 (2009).
25. Krogstrup, P. *et al.* Impact of the liquid phase shape on the structure of III–V nanowires. *Phys. Rev. Lett.* **106,** 125505 (2011).
26. Munshi, A. M. *et al.* Crystal phase engineering in self-catalyzed GaAs and GaAs/GaAsSb nanowires grown on Si(111). *J. Cryst. Growth* **372,** 163–169 (2013).
27. Cirlin, G. E. *et al.* Self-catalyzed, pure zincblende GaAs nanowires grown on Si(111) by molecular beam epitaxy. *Phys. Rev. B* **82,** 035302 (2010).
28. Spirkoska, D. *et al.* Structural and optical properties of high quality zinc-blende/wurtzite GaAs nanowire heterostructures. *Phys Rev B* **80,** 245325 (2009).
29. Ross, F. M. Controlling nanowire structures through real time growth studies. *Rep. Prog. Phys.* **73,** 114501 (2010).
30. Chou, Y.-C. *et al.* Atomic-scale variability and control of III–V nanowire growth kinetics. *Science* **343,** 281–284 (2014).
31. Wen, C. Y. *et al.* Periodically changing morphology of the growth interface in Si, Ge, and GaP nanowires. *Phys. Rev. Lett.* **107,** 025503 (2011).
32. Prince, A. A., Raynor, G. V. & Evans, D. S. *Phase Diagrams of Ternary Gold Alloys* 123–132 (Institute of Metals, 1990).
33. Wen, C. Y., Reuter, M. C., Tersoff, J., Stach, E. A. & Ross, F. M. Structure, growth kinetics, and ledge flow during vapour–solid–solid growth of copper-catalyzed silicon nanowires. *Nano Lett.* **10,** 514–519 (2010).
34. Oh, S. H. *et al.* Oscillatory mass transport in vapor-liquid-solid growth of sapphire nanowires. *Science* **330,** 489–493 (2010).
35. Gamalski, A. D., Ducati, C. & Hofmann, S. Cyclic supersaturation and triple phase boundary dynamics in germanium nanowire growth. *J. Phys. Chem. C* **115,** 4413–4417 (2011).
36. Hilner, E. *et al.* Au wetting and nanoparticle stability on GaAs(111)B. *Appl. Phys. Lett.* **89,** 251912 (2006).
37. Schwarz, K. W. & Tersoff, J. Elementary processes in nanowire growth. *Nano Lett.* **11,** 316–320 (2011).
38. Yu, X. *et al.* Evidence for structural phase transitions induced by the triple phase line shift in self-catalyzed GaAs nanowires. *Nano Lett.* **12,** 5436–5442 (2012).
39. Krogstrup, P. *et al. In-situ* x-ray characterization of wurtzite formation in GaAs nanowires. *Appl. Phys. Lett.* **100,** 093103 (2012).
40. Takahasi, M., Kozu, M., Sasaki, T. & Hu, W. Mechanisms determining the structure of gold-catalyzed GaAs nanowires studied by in situ X-ray diffraction. *Cryst. Growth Des.* **15,** 4979–4985 (2015).

41. Johansson, J., Zanolli, Z. & Dick, K. A. Polytype attainability in III–V semiconductor nanowires. *Cryst. Growth Des.* **16,** 371–379 (2016).
42. Schmidt, V., Senz, S. & Gösele, U. The shape of epitaxially grown silicon nanowires and the influence of line tension. *Appl. Phys. A* **80,** 445–450 (2005).
43. Jiang, N. *et al.* Understanding the true shape of Au-catalyzed GaAs nanowires. *Nano Lett.* **14,** 5865–5872 (2014).
44. Hillerich, K. *et al.* Strategies to control morphology in hybrid group III–V/group IV heterostructure nanowires. *Nano Lett.* **13,** 903–908 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements D.J., S.L. and K.A.D. acknowledge financial support from the Knut and Alice Wallenberg Foundation (KAW), the Swedish Research Council (VR) and the Nanometer Structure Consortium at Lund University (nmC@LU). F.P. and S.H. acknowledge support from ERC Grant 279342: InSituNANO. We acknowledge A. Ellis for technical support.

Author Contributions D.J. and F.P. performed experiments and data analysis, J.T. developed the model, M.C.R. developed the UHVTEM technique, S.L. provided growth expertise, and K.A.D., S.H. and F.M.R. designed the experiments and coordinated the analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.J. (Daniel.jacobsson@ftf.lth.se) or F.M.R. (fmross@us.ibm.com).

METHODS

The GaAs nanowires imaged in this study were grown on Si(111). The substrates were cut from a Si(111) wafer into strips 3 mm × 350 μm 500 μm, small enough to fit directly into the TEM heating holder. The strips were, however, too small to be handled in the GaAs growth system, so they were stacked in arrays, parallel to each other with the polished surface facing upwards, and mounted on a larger Si wafer. At Lund University, Au aerosol particles with diameters of 30 nm, 50 nm and 70 nm were deposited onto the arrays of strips using a size-selected aerosol source at a total density of about 1 particle per μm^2. Then, GaAs nanowires of the order of 500 nm in length were grown on the arrays using standard metal–organic vapour phase epitaxy in an Epiquip system, operating at 100 mbar with AsH_3 and TMGa as precursor gases and H_2 as carrier gas. After growth, the arrays were glued to sample boxes using a small piece of SEM-type double-sided carbon tape, the sample boxes were placed in a plastic bag, which was vacuum sealed, and then the bag was sent through air to the UHVTEM at IBM. The individual strips were separated and each sample was degassed in UHV by resistive heating below 100 °C for 30 min, flowing a direct current through the Si strip. The heating current required for a temperature of around 300 °C was then determined in a separate UHV chamber using an infrared pyrometer. All of the strips had a similar temperature–current calibration, so it was possible to estimate the current required to heat the sample to 500 °C or 550 °C. The sample was transferred to the UHVTEM column to check that the nanowires and Au catalysts were still present after this process. Finally, TMGa was flowed to a chosen pressure of around 5×10^{-8} Torr as measured using a mass spectrometer, AsH_3 was flowed to a chosen pressure of around 2×10^{-5} Torr as measured on the column ion gauge, the nanowires were heated to 500–550 °C and GaAs was grown at the nanowire tips. The crystal phase was generally controlled using AsH_3 pressure, which was easier to measure and faster to change than TMGa. After experiments on one sample were completed, the full current–temperature calibration curve was obtained for that sample. The reason for this calibration procedure was to prevent any damage (for example, etching) of the wires by overheating before the growth experiment began. Owing to drift of the temperature on continued heating, we estimate the temperature accuracy to be ±20 °C. All observations of crystal switching occurred between 500 °C and 570 °C. Approaching 500 °C, the temperature range over which switching occurred became narrower and ZB grew for all accessible pressures. Above 600 °C, the nanowires etched slowly at the Au/GaAs interface, presumably owing to the low group V pressure.

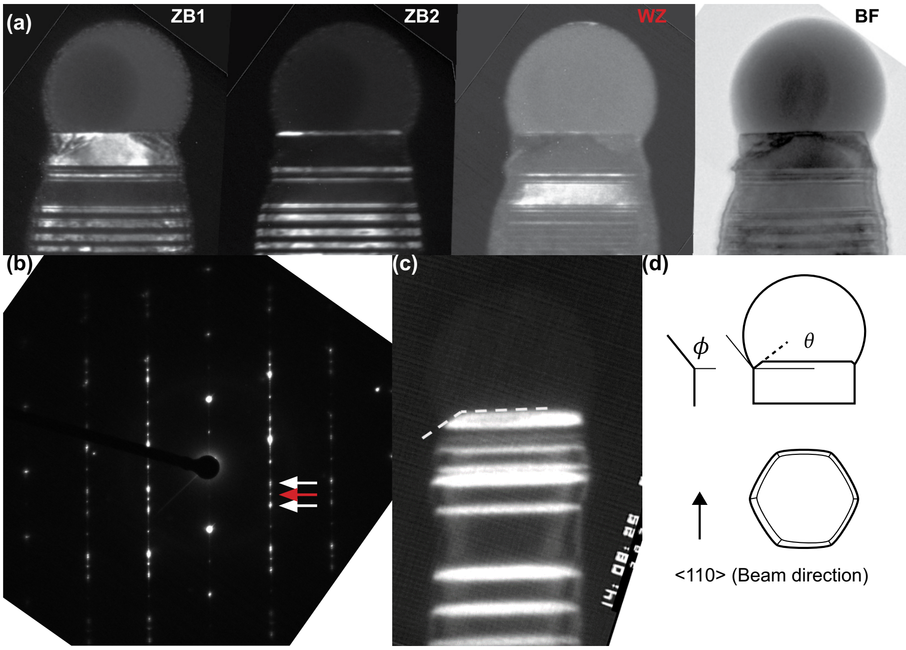
This growth *in situ* within the TEM is somewhat different from standard MOVPE. Even though the conventional MOVPE precursor gases are used, there is no H_2 carrier gas during growth within the TEM, as is typically used during MOVPE growth of GaAs. In addition, the absolute pressures of the two precursor species are lower than typical precursor partial pressures used in MOVPE. The lower partial pressures can alone account for the low growth rates observed here compared to those observed in MOVPE. To compare the effects of the V/III ratio, we need to consider the possible differences between the two methods in more detail.

The growth in this study was observed to always be group-V limited, as seen in Extended Data Fig. 2c. This is in contrast to standard MOVPE, where high group-V flows and group-III-limited regimes are typically used. Instead, one could argue that the *in situ* TEM conditions are more similar or relevant to chemical beam epitaxy (CBE), or possibly molecular beam epitaxy (MBE), than to MOVPE. The group-V-limited regime is also highly relevant to catalyst-free growth from Ga droplets. One exception for MOVPE is a recent study[9] exploring group-V-limited regimes in a standard MOVPE reactor to grow WZ and ZB GaAs. In this work, it was shown that at low enough group-V flow to yield group-V-limited growth, WZ grows at 'high' V/III ratio, whereas ZB grows when the V/III ratio is lowered. This result is in contrast to the more well-known behaviour in the group-III-limited regime, in which ZB forms at high V/III ratio; however, the result of ref. 9 is entirely in agreement with the results of this study. Where the present work and that presented in ref. 9 differ is in the absolute magnitude of the V/III ratio: here, V/III ratios of 100 or more yielded group-V-limited nanowire growth; in ref. 9, the group-V-limited regime occurred at V/III ratios of less than 2.

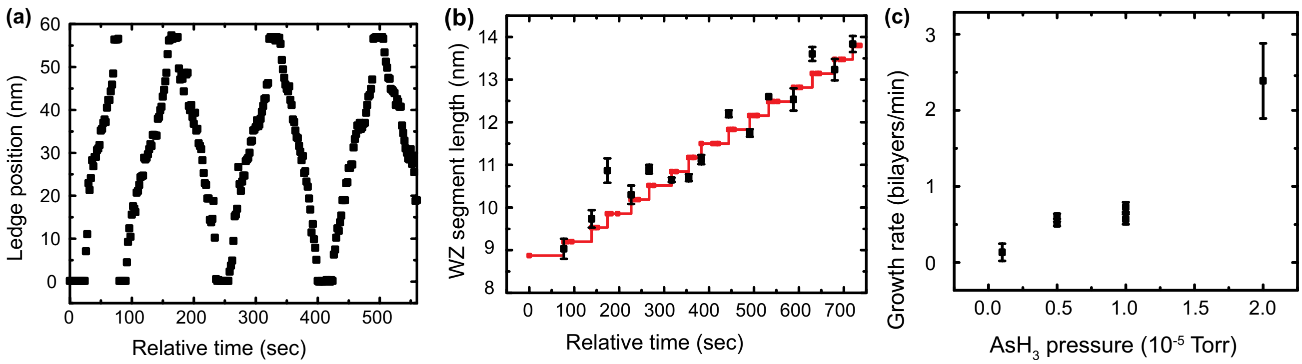
This comparison suggests that the effective As pressure at the growth front is substantially lower in the TEM, relative to the Ga pressure, than in typical MOVPE. To understand this, we note that AsH_3 pyrolysis in GaAs growth is generally considered to proceed heterogeneously on GaAs surfaces, without interaction with the carrier gas[45]. This pyrolysis starts with adsorption of AsH_3 onto the surface, followed by sequential dissociation of H atoms one by one, eventually leaving atomic As adsorbed on the surface[46]. When AsH_3 is combined with TMGa, however, the two species decompose together, simultaneously, via adduct formation on the surface; this decomposition pathway does not involve hydrogen and is more efficient than the decomposition of either species alone[45]. That the species decompose primarily on the surface is an important clue to the relatively inefficient supply of As in the UHVTEM. First, the nanowires are grown on Si substrates rather than on GaAs; although there is also ample GaAs surface on the pre-grown nanowire stubs, this surface is clearly different from the typical GaAs substrates used for MOVPE nanowire growth. Second, the surfaces are likely to be passivated with hydrogen when growth occurs in a H_2 atmosphere; the absence of H_2 here could affect the supply in a number of ways, changing, for example, the decomposition process and precursor surface diffusion. Finally, the decomposition process relies on the desorption of gas-phase As species. This adsorption process naturally depends on the partial pressure; because As has a substantially higher vapour pressure than Ga, the adsorption process of As will be reduced to a greater extent by the lower partial pressure.

Other minor differences between growth in the TEM and growth in a reactor are expected, owing to the experimental set-up. When using needle valves rather than standard mass-flow controllers to control the precursor flows, the experimental parameters are less accurately controlled than they are in dedicated epitaxy growth systems. At the high V/III ratios used (V/III > 100), the TMGa partial pressure is much lower than that of AsH_3. A gauge reading the total pressure close to the sample is used to monitor the AsH_3 flow and provide fast feedback on the AsH_3 pressure at the sample. To monitor TMGa pressure, a mass spectrometer is used, with a controlled, steady pressure of TMGa set at the start of the experiment and generally held constant. The mass spectrometer is continuously used during the experiments to monitor any drift in the TMGa pressure.

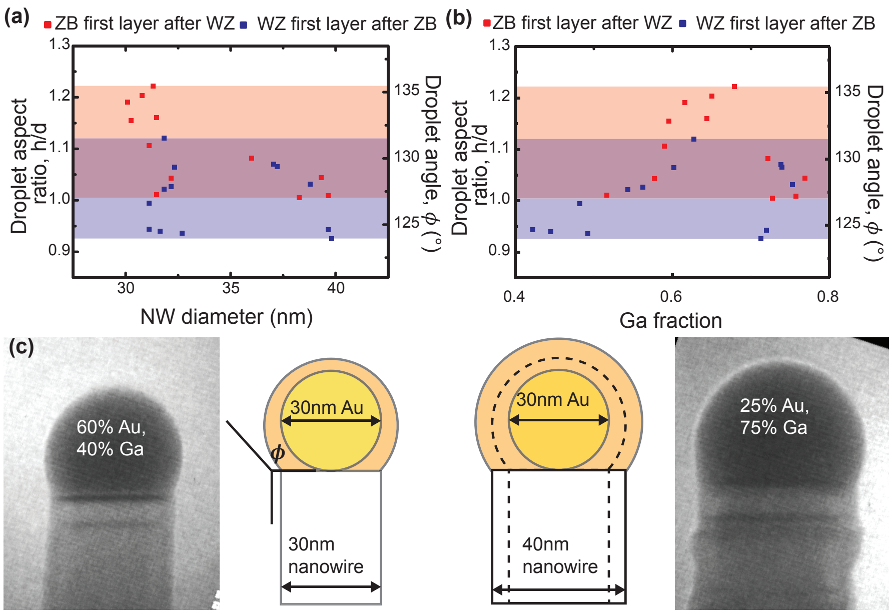
45. Larsen, C. A., Buchan, N. I. & Stringfellow, G. B. Reaction mechanisms in the organometallic vapor phase epitaxial growth of GaAs. *Appl. Phys. Lett.* **52,** 480–482 (1988).
46. Tamaru, K. The decomposition of arsine. *J. Phys. Chem.* **59,** 777–780 (1955).



Extended Data Figure 1 | Distinguishing crystal phases using dark-field imaging. **a**, Dark-field images recorded using three spots in the diffraction pattern showing how WZ and the two variants of ZB (ZB1 and ZB2) are distinguished in the ⟨110⟩ direction. The right panel shows a bright-field (BF) image for comparison. **b**, Diffraction pattern with the three spots indicated. These are post-growth images recorded in a JEOL 3000 TEM. **c**, Image recorded during growth, also in the ⟨110⟩ direction, showing the dark-field contrast and the measured angle of the truncated corner. **d**, A schematic showing the hexagonal cross-section as seen from the side (beam direction) and from above. Seen from the side, the droplet contact angle ϕ and corner-facet angle θ are shown. Seen from above, the electron beam direction is shown.



Extended Data Figure 2 | Kinetics of WZ growth. a, Step-flow kinetics measured for the WZ nanowire shown in Supplementary Video 1 and Fig. 1a. **b**, Length of a WZ nanowire versus the number of step-flow events. The average gradient of the graph confirms that each step is 0.33 nm in height, that is, a WZ(0001) bilayer. The red line shows the calculated length, assuming that each new bilayer (red squares) adds 0.33 nm in length. The black data are measured lengths, with error bars defined by the standard deviation of each subset (length measured in 3–4 different movie frames for each data point). **c**, Growth rate versus AsH_3 pressure, estimated by measuring the increase in the length of the nanowire shown in Fig. 2 during growth intervals at different AsH_3 pressures, and by fitting a linear function to the length versus time plot. Error bars are defined by the standard error of the fits. The growth temperature was 550 °C and TMGa pressure was constant at 3.5×10^{-8} Torr.



Extended Data Figure 3 | Droplet angle at the transition does not change with Au composition in the droplet. a–**c**, During growth, the crystal structure of a nanowire was switched back and forth several times by changing the AsH_3 pressure. Data shows the measured angles and absolute droplet volumes at which the switch from WZ to ZB (red) and ZB to WZ (blue) occurred (**a**, **b**). During this time the wire also grew radially (**c**). **a**, Switch angles versus nanowire diameter. Shading indicates the range of observed angles at which WZ switches to ZB (red) and ZB to WZ (blue). Some hysteresis in switching is visible, perhaps because the droplet angle continues to change in the time before the switched layer grows. The data are scattered, but there is no strong dependence of angle on diameter (especially for the blue data points). For example, ZB switches to WZ at angles between 123° and 132° at both large and small diameter, and WZ switches to ZB at angles between 127° and 136°. **b**, Switch angle versus inferred composition, as calculated from the measured *h*/*d* ratio assuming that the amount of Au does not change. The data are scattered, but there is no strong dependence of angle on composition. For example, for the first few data points, for which the nanowire had a diameter of 30 mn, ZB switched to WZ for a Ga fraction of less than 60%. For subsequent data points, for which the nanowire had a diameter of 40 nm, the switch did not occur until the composition had a Ga fraction of about 75%. The fact that droplet angles are similar despite the change in diameter suggests that droplet geometry controls the switch, whereas droplet volume and composition do not appear to be important.

doi:10.1038/nature17181

Lens regeneration using endogenous stem cells with gain of visual function

Haotian Lin[1]*, Hong Ouyang[1]*, Jie Zhu[2]*, Shan Huang[1]*, Zhenzhen Liu[1], Shuyi Chen[1], Guiqun Cao[3], Gen Li[3,4], Robert A. J. Signer[5], Yanxin Xu[3,6], Christopher Chung[2], Ying Zhang[7], Danni Lin[2], Sherrina Patel[2], Frances Wu[2], Huimin Cai[3,4], Jiayi Hou[8], Cindy Wen[2], Maryam Jafari[2], Xialin Liu[1], Lixia Luo[1], Jin Zhu[2], Austin Qiu[2], Rui Hou[4], Baoxin Chen[1], Jiangna Chen[1], David Granet[2], Christopher Heichel[2], Fu Shang[1], Xuri Li[1], Michal Krawczyk[2], Dorota Skowronska-Krawczyk[2], Yujuan Wang[1], William Shi[2], Daniel Chen[2], Zheng Zhong[1,2], Sheng Zhong[2], Liangfang Zhang[2], Shaochen Chen[2], Sean J. Morrison[5], Richard L. Maas[7], Kang Zhang[1,2,3,9] & Yizhi Liu[1]

The repair and regeneration of tissues using endogenous stem cells represents an ultimate goal in regenerative medicine. To our knowledge, human lens regeneration has not yet been demonstrated. Currently, the only treatment for cataracts, the leading cause of blindness worldwide, is to extract the cataractous lens and implant an artificial intraocular lens. However, this procedure poses notable risks of complications. Here we isolate lens epithelial stem/progenitor cells (LECs) in mammals and show that *Pax6* and *Bmi1* are required for LEC renewal. We design a surgical method of cataract removal that preserves endogenous LECs and achieves functional lens regeneration in rabbits and macaques, as well as in human infants with cataracts. Our method differs conceptually from current practice, as it preserves endogenous LECs and their natural environment maximally, and regenerates lenses with visual function. Our approach demonstrates a novel treatment strategy for cataracts and provides a new paradigm for tissue regeneration using endogenous stem cells.

Stem-cell therapy holds great promise in regenerative medicine. Much attention has been focused on pluripotent stem cells and the use of their derivatives for therapeutic purposes. However, several uncertainties, including tumorigenicity and immune rejection, have hindered their clinical application. An attractive alternative is to harness the potential of endogenous stem/progenitor cells for direct use in repair and regeneration. In the case of the ocular lens, regeneration has been reported in lower vertebrates[1,2]. In mammals, such as rabbits, removal of the original lens content results in the proliferation of residual lens epithelial stem/progenitor cells (LECs) and the generation of a limited amount of lens fibres[3,4]. In humans, varying degrees of disorganized regrowth of doughnut-like lens tissues have been observed after congenital cataract removal in infants (Extended Data Fig. 1a, b). However, the underlying mechanism for these observations remains elusive, and the successful regeneration of a complete mammalian lens with biological function has yet to be achieved.

Cataracts are the leading cause of blindness in the world[5]. The visual axis, defined as the normal passage of light into the eye, may undergo visual axis opacification (VAO) owing to the cataractous lens or the postoperative disorganized growth of remaining LECs, leading to vision loss[6]. The current standard of care in congenital cataracts involves surgical removal of the cataractous lens with a large central capsulorhexis opening and implantation of an artificial intraocular lens (IOL) to replace the missing refractive media. Although artificial IOLs are widely used in paediatric cataract surgery, they are limited by complications[7,8], and most paediatric patients continue to require some form of refractive correction such as eyeglasses after cataract surgery[9]. Furthermore, IOLs are controversial in patients younger than two years as they have not been shown to prevent strabismus or amblyopia, and normal lens refractive power is not yet fully developed at this age[10,11].

To test the feasibility of *in situ* lens regeneration, we performed *in vitro* studies on $PAX6^{+}/SOX2^{+}$ LECs and identified BMI-1 as an essential factor for maintaining a LEC pool in mammalian eyes by conditional knockout experiments. We also investigated the ability of LECs to differentiate into lens fibre cells *in vitro*. We then performed *in vivo* animal studies by establishing a new minimally invasive capsulorhexis surgery method that differs conceptually from current practice in extracting the cataractous lens through a small wound opening, while preserving lens capsule integrity and therefore LECs as well. Using this method, we investigated lens regeneration in rabbits and macaques and conducted a clinical trial in human infants. Functional lens regeneration was observed not only in rabbits and macaques, but also in human patients with congenital cataracts. Therefore, our new study provides a novel approach to lens regeneration using endogenous stem cells, and results in improved outcomes.

Essential role of LECs in lens regeneration

In the mature lens, LECs cover the anterior surface of the lens and begin to differentiate into lens fibres at the equator (Extended Data Fig. 2a). Sustained self-renewal and protective capacities against external injury and oxidative damage are among the most significant functions of LECs[12]. To assess the regenerative ability of LECs, we used bromodeoxyuridine (BrdU) labelling to identify proliferating LECs from human donor lenses. We quantified $BrdU^{+}$ LECs in 8-month-old, 30-year-old, and 40-year-old donors and found that the number of proliferating cells decreased with age (Extended Data Fig. 2b, c). However, upon surgical removal of the entire lens contents with preservation of the empty

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China. [2]Shiley Eye Institute, Institute for Engineering in Medicine, Institute for Genomic Medicine, University of California, San Diego, La Jolla, California 92093, USA. [3]Molecular Medicine Research Center, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Sichuan 610041, China. [4]Guangzhou KangRui Biological Pharmaceutical Technology Company, Guangzhou 510005, China. [5]Howard Hughes Medical Institute, Children's Research Institute, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [6]Department of Ophthalmology, West China Hospital, Sichuan University, Sichuan 610041, China. [7]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [8]Clinical and Translational Research Institute, University of California, San Diego, La Jolla, California 92093, USA. [9]Veterans Administration Healthcare System, San Diego, California 92093, USA.
*These authors contributed equally to this work.

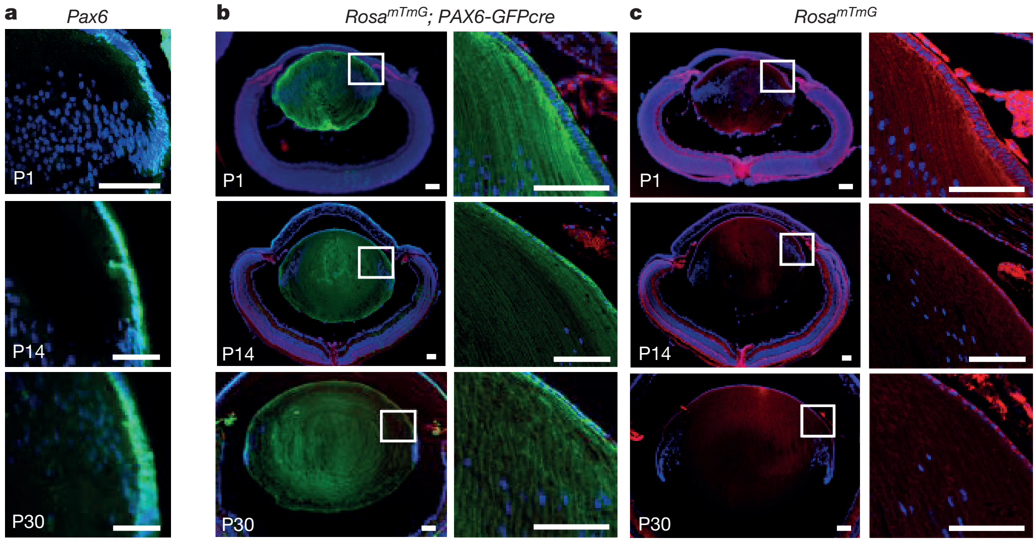


Figure 1 | Lineage tracing of Pax6$^+$ LECs in mice. a, *Pax6*-directed GFP was expressed in mouse LEC nuclei at post-natal days P1, P14 and P30; a sagittal section of a *P0-3.9-GFPcre* mouse lens is shown. Blue and green represent DAPI and GFP, respectively. **b**, Lineage tracing of Pax6$^+$ LECs in *ROSA^{mTmG}*; *P0-3.9-GFPcre* mice at P1, P14 and P30 reveals that lens fibre cells express membrane GFP fluorescence; hence, PAX6$^+$ LECs were able to generate lens fibre cells. **c**, As an additional control, the *ROSA^{mTmG}* allele alone exhibits Tomato (red) staining at sites of non-recombination. All scale bars, 100 µm.

capsular bag scaffold, the number of BrdU$^+$ cells increased by 11-fold ($P < 0.05$, Extended Data Fig. 2d, e), suggesting a strong regenerative capacity of human LECs after injury.

Pax6 plays a central role in eye development as well as in lens induction. After birth, *Pax6* maintains a high level of expression in the lens epithelium, particularly at the germinative zone (Fig. 1a). To determine whether Pax6$^+$ LECs can contribute to lens fibre cell formation, we performed lineage-tracing experiments in mice by crossing a *Pax6* lens ectoderm enhancer-driven Cre deleter mouse strain (*P0-3.9-GFPcre*) with the *ROSA^{mTmG}* membrane-bound GFP reporter strain. We observed intense membrane GFP$^+$ cells throughout the entire lens of *ROSA^{mTmG}*; *Pax6P0-3.9-GFPcre* mice at P1, P14, and P30; in contrast, the *P0-3.9-GFPCre* allele alone yielded only nuclear GFP expression in LECs detectable by anti-GFP antibody staining (Fig. 1a, b). These results indicate that Pax6$^+$ LECs from the embryonic or adult lens can contribute to the post-natal replacement of mouse lens fibre cells.

We next isolated and expanded rabbit LECs from neonatal lens capsules (see Methods). These LECs showed a cobble-stone-like epithelial morphology with highly positive staining for LECs markers Pax6 and Sox2, and could be passaged over time (Fig. 2a). Upon differentiation, these LECs formed transparent three-dimensional convex lens-like structures, defined as lentoid bodies (Fig. 2b, c), which possess significant refractive power (Fig. 2c). Immunostaining and western blot analysis showed that lentoid bodies expressed mature lens-fibre-specific genes, including those encoding αA-, β-, and γ-crystallins (Fig. 2b, c).

Loss of LEC homeostasis leads to cataracts

To further investigate the LEC pool and its role in the maintenance of lens function, we studied BMI-1, a member of the Polycomb-group family. BMI-1 is known to promote the maintenance and self-renewal of stem cells in multiple post-natal tissues[13–15] and is expressed in both the murine lens germinative zone and in cultured human fetal LECs (Extended Data Fig. 2f, g and 3a). Knockdown of *BMI1* in human LECs led to significantly decreased LECs proliferation *in vitro* (Extended Data Fig. 4a), without affecting expression of key genes in LECs or lens fibre cells (Extended Data Fig. 4b). To directly test the effects of conditional deletion of *Bmi1* on LEC proliferation, we administered BrdU to 2-, 7-, and 12-month-old *Nestin-cre;Bmi1$^{fl/fl}$* mice and *Bmi1$^{fl/fl}$* littermate controls. After a 20-h pulse, there was no significant difference in the percentage of BrdU$^+$ LECs in 2-month-old *Nestin-cre;Bmi1$^{fl/fl}$* mice and *Bmi1$^{fl/fl}$* controls. However, there was a significant reduction in the percentage of BrdU$^+$ LECs in 7- and 12-month-old *Nestin-cre;Bmi1$^{fl/fl}$* eyes compared to controls (Extended Data Fig. 3b, $P < 0.05$).

We next investigated the mRNA expression levels of *Bmi1*, *Sox2 and Ki67* in Pax6$^+$ LECs at the anterior capsule in *Pax6P0-3.9-GFPcre* mouse lens. Compared with Pax6$^-$ (GFP-negative) LECs, Pax6$^+$ (GFP-positive) LECs located at the germinative zone had higher expression levels of *Bmi1*, *Sox2 and Ki67* (Extended Data Fig. 5). Moreover, conditional deletion of *Bmi1* led to a dramatic decrease in the number of Pax6$^+$/Sox2$^+$ LECs in ageing *Nestin-cre;Bmi1$^{fl/fl}$* mice (Extended Data Fig. 3a, $P < 0.001$). Additionally, the lenses of ageing *Nestin-cre;Bmi1$^{fl/fl}$* mice became progressively opaque, suggesting cataract formation.

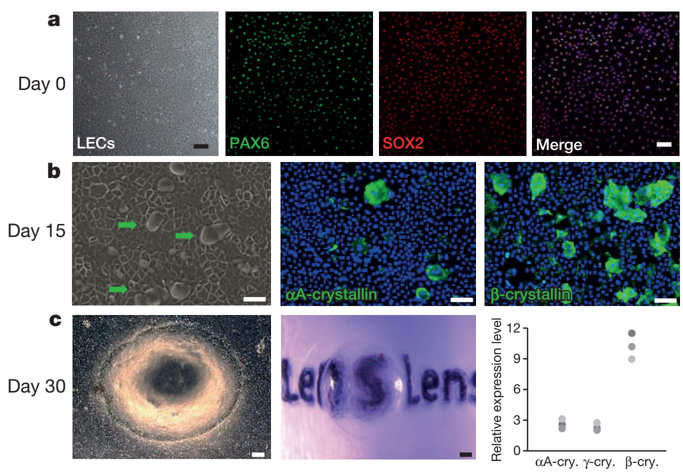


Figure 2 | Characterization and differentiation of rabbit LECs. a, LECs were positive for PAX6 (green) and SOX2 (red). **b**, Lentoid formation (green arrows) with positive αA-crystallin and β-crystallin staining on day 15 of LEC differentiation. **c**, Left panel, phase-contrast photograph of a lentoid body on day 30; middle panel, a lentoid body demonstrating magnifying properties; right panels, photograph from western blot analysis and quantification showing a dramatic increase in expression relative to pre-differentiation expression of mature lens fibre markers αA-crystallin (2.6, 3.1, 2.2), β-crystallin (11.51, 9.0, 10.2) and γ-crystallin (2.2, 2.0, 2.8) (numbers in parentheses represent fold change after differentiation). $n = 3$ biological replicates. All scale bars, 100 µm.

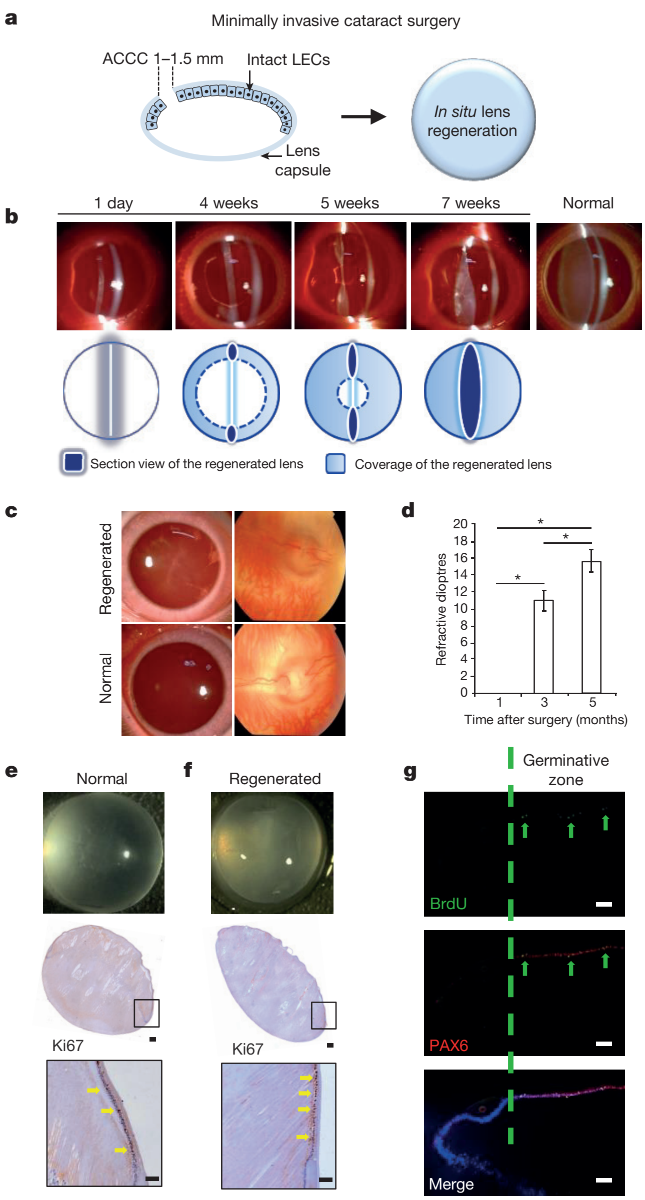


Figure 3 | Lens regeneration in rabbits. a, New minimally invasive surgical method. The capsulorhexis size was decreased to 1.0–1.5 mm in diameter, resulting in a reduced wound area of 1.2 mm^2, and moved to the periphery of the lens. **b**, Slit-lamp microscopy showing the progress of lens regeneration after minimally invasive surgery in a rabbit eye. **c**, Fundus examination of rabbit eyes 7 weeks post-surgery demonstrated a clearly visible retina. Normal healthy lens shown for comparison. **d**, Measurements of refractive dioptres in rabbit eyes at different time points post-surgery (M, month; D, dioptres). Refractive dioptres of the eyes increased with time after surgery, demonstrating the functionality of the regenerated lenses (ANOVA, $*P < 0.01$). The refractive power immediately after surgery was defined as zero, 1 month = 0.0 dioptre, 3 months = 11.0 ± 0.8 dioptres and 5 months = 15.8 ± 2.2 dioptres, $n = 6$ at each time point, data shown as means ± s.d. **e**, **f**, Ki67 staining in the germinative zone of normal rabbit lens (**e**) and regenerated rabbit lens 7 weeks post-surgery (**f**). Lower panels show higher magnification. **g**, PAX6 (red) and BrdU (green) staining at the germinative zone of regenerated rabbit lens 7 weeks post-surgery. Scale bars, 100 µm.

To test this hypothesis, we administered tropicamide drops to the eyes of 2-, 7-, and 12-month-old *Nestin-cre;Bmi1*$^{fl/fl}$ mice and *Bmi1*$^{fl/fl}$ littermate controls to dilate the pupils (Extended Data Fig. 3c, d). Eyes of 2-month-old *Nestin-cre;Bmi1*$^{fl/fl}$ mice ($n = 3$) were indistinguishable from those of age-matched controls ($n = 4$). However, 100% of the 7-month-old ($n = 5$) and 12-month-old ($n = 7$) *Nestin-cre;Bmi1*$^{fl/fl}$ mice had bilateral cataracts, while none of the age-matched *Bmi1*$^{fl/fl}$ controls ($n = 3$, 7-month-old; $n = 5$, 12-month-old) developed cataracts. Moreover, haematoxylin and eosin-stained sections revealed the presence of cataracts in the 7- and 12-month-old *Nestin-cre;Bmi1*$^{fl/fl}$ mice (Extended Data Fig. 3d). These data suggest that *Bmi1* loss of function disrupted LEC proliferation, thereby depleting the LEC pool and promoting cataract formation.

LEC preservation and lens regeneration

The current capsulorhexis method performed in paediatric cataract surgery involves making a large 6-mm-diameter opening at the centre of the anterior capsule, resulting in a large wound area and destruction of large numbers of LECs (Extended Data Fig. 1c). To overcome these limitations and to facilitate lens regeneration, we established a new capsulorhexis method. This new method has two advantages: (1) it reduces the size of the wound considerably; and (2) it moves the capsulorhexis opening from the central visual axis to the periphery. Thus, application of this procedure led to improved visual axis transparency and preservation of LECs with regenerative potential (Fig. 3a).

We next investigated lens regeneration in rabbit eyes *in vivo*. We used our new minimally invasive capsulorhexis technique to preserve endogenous LECs while removing the native lens (Extended Data Fig. 6). One day after surgery, slit-lamp microscopy showed that the anterior and posterior capsules were adherent (Fig. 3b). Four to five weeks after surgery, the regenerating lens tissue grew from the periphery of the capsular bag towards the centre in a curvilinear symmetrical pattern (Fig. 3b). Seven weeks after surgery, the regenerating lens tissue formed a transparent biconvex lens along the anterior–posterior axis with a clear view of the posterior segment and retina (Fig. 3b, c), comparable to a normal healthy lens (Fig. 3c). The refractive power of the regenerated lenses after surgery was evaluated and found to have increased to an average of 15.6 dioptres from the first to the fifth month after surgery, a value comparable to that of a normal lens[16] (Fig. 3d, $P < 0.01$).

In contrast to the epithelial cells and premature lens fibre cells located at the lens equator, the LECs in the germinative zone of regenerated lenses showed intense proliferative activity 7 weeks post-surgery, as evidenced by both Ki67 and BrdU labelling (Fig. 3e–g). Notably, some PAX6$^+$ LECs co-labelled with BrdU, demonstrating their proliferative potential (Fig. 3g). These LECs lost PAX6 expression concomitant with the initiation of differentiation and subsequent migration from the lens equator.

One day post-surgery, histological examination revealed that a monolayer of LECs remained intact (Extended Data Fig. 7a). Four days post-surgery, LECs migrated onto the posterior capsule from the periphery towards the centre in a curvilinear 360° fashion with a single layer of epithelium on the posterior capsule (Extended Data Fig. 7a). Seven days post-surgery, LECs on the posterior capsule began to elongate, and their nuclei were positioned anteriorly away from the posterior capsule (Extended Data Fig. 7a). Twenty-eight days post-surgery, a structure with lens fibres and an extruded nucleus was observed (Extended Data Fig. 7b). At week 7 after surgery, the regenerated lens fibres elongated along the anterior–posterior axis and grew to cover the entire posterior capsular area, forming a lens with a double-convex shape (Extended Data Fig. 7c).

We next investigated lens regeneration in macaques 1–3 months of age (approximately equivalent to human infants 4–12 months old), using a similar minimally invasive surgical technique. From postoperative days 1 to 3, no signs of inflammation or other undesired side-effects were seen. Two to three months post-surgery, regenerating lens tissue had grown from the periphery towards the centre in a curvilinear pattern (Fig. 4a). Five months post-surgery, a biconvex lens with a transparent visual axis had formed (Fig. 4a, b). Fundus examination 7 weeks after surgery showed a clear view of the retina, comparable to the view of the retina seen through a normal healthy lens. No undesired complications, such as macular oedema, retinal detachment, or endophthalmitis were observed.

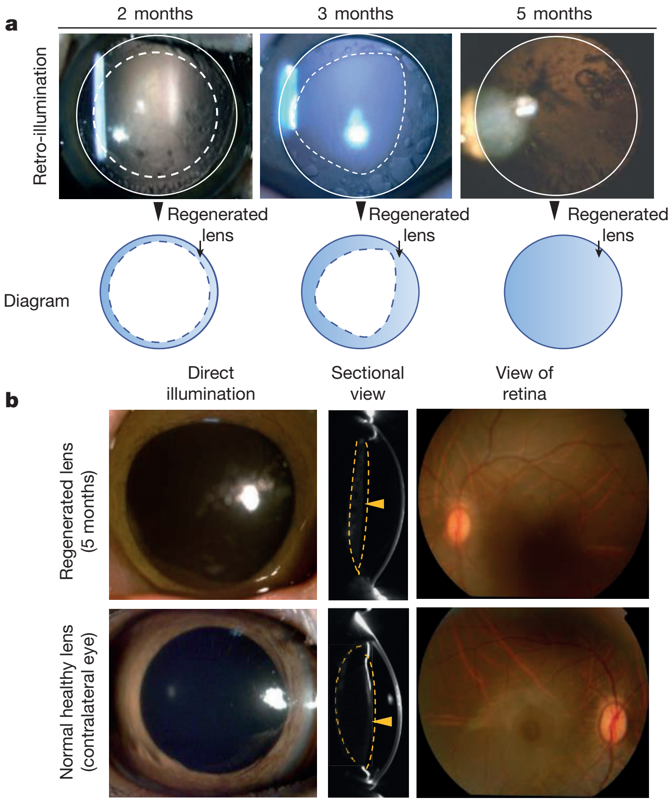


Figure 4 | Lens regeneration in macaque models after minimally invasive surgery. a, Slit-lamp microscopy showed regenerating lens tissue grew from the peripheral to the central lens in a circular symmetrical pattern 2–3 months after surgery, reaching the centre at 5 months post-surgery. Five months after surgery, direct illumination showed that the visual axis remained translucent. **b**, Pentacam cross-sectional scanning showed formation of a biconvex structure 5 months after surgery (yellow arrowheads). Direct illumination and fundus photography showed that the visual axis remained transparent and the retina was clearly visible ($n=6$).

Lens regeneration in human infants

Cataracts are a major cause of vision loss in human infants[17]. Currently, the most commonly practiced surgical procedure involves removal of the cloudy lens through a large anterior continuous curvilinear capsulorhexis (ACCC), combined with either posterior laser capsulotomy or posterior continuous curvilinear capsulorhexis (PCCC) and anterior vitrectomy (Extended Data Fig. 1), which is followed by artificial lens implantation or postoperative aphakic eyeglasses or contact lenses[18]. However, complications such as VAO often occur. Moreover, difficulty with refractive correction of developing eyes, secondary glaucoma, and surgery-related complications can lead to a poor outcome[19]. We conducted a clinical trial in paediatric cataract patients up to two years of age to investigate whether lenses could be regenerated in humans using minimally invasive surgery.

Twelve paediatric cataract patients (24 eyes) underwent minimally invasive surgery to promote lens regeneration, while 25 paediatric cataract patients (50 eyes) in the control group received the current standard-of-care treatment that left them aphakic. Using slit-lamp microscopy, we were able to dynamically observe the process of *in vivo* lens regeneration postoperatively. The capsular openings healed within one month after minimally invasive surgery. Three months post-surgery, a regenerated transparent biconvex lens structure had formed (Extended Data Fig. 8c, d). No significant VAO or other complications were observed at 6 months post-surgery (Table 1).

All of the eyes gained visual function when the capsular bag was refilled with a regenerated lens of relatively uniform density. A clear view of the fundus was observed in all cases with successful lens regeneration (Extended Data Fig. 8c, d). The average central thickness of the regenerated lenses increased significantly after surgery and was comparable to a native lens at 8 months post-surgery (Fig. 5a, $P<0.01$). We also used retinoscopy and ophthalmoscopy to evaluate the function of the regenerated lenses and found that from the first week to 8 months post-surgery, the refractive power increased significantly (Fig. 5b, $n=24$, $P<0.01$).

The accommodative ability of the regenerated lenses was evaluated 8 months after surgery using an open-field autorefractor (see Methods). The mean accommodative response increased to 2.5 dioptres in

Table 1 | Comparison of lens regeneration and complications in infants who received the new surgical treatment versus the current treatment

	Current treatment		New treatment			
Total patients	25		12			
Total eyes	50		24			
Regenerated lens structure	0		24			
Healing and closure of capsular openings	0		24			
	Current treatment decimal acuity (logMAR)		New treatment decimal acuity (logMAR)			
	OD	OS	OD	OS		
Baseline	0.008 (0.001)	0.008 (0.001)	0.008 (0)	0.008 (0.001)		
1 week	0.03 (0.009)	0.03 (0.010)	0.03 (0.007)	0.03 (0.008)		
3 months (before laser)	0.02 (0.017)	0.02 (0.022)	0.05 (0.014)	0.05 (0.017)		
3 months (after laser)	0.05 (0.013)	0.05 (0.018)	–	–		
6 months	0.11 (0.034)	0.11 (0.025)	0.10 (0.038)	0.11 (0.027)		
	Current treatment		New treatment		Mean difference (95% CI)	*P* value
Overall complication rate	46 (0.92)		4 (0.17)		0.75 (0.57, 0.95)	<0.001
Corneal oedema	15 (0.30)		2 (0.08)		0.22 (0.02, 0.42)	0.04
Anterior chamber inflammation	37 (0.74)		4 (0.17)		0.57 (0.35, 0.80)	<0.001
Macular oedema	3 (0.06)		0		0.06 (−0.04, 0.16)	0.22
Endophthalmitis	0		0		–	–
Retinal detachment	0		0		–	–
Ocular hypertension	9 (0.18)		0		0.18 (0.04, 0.32)	0.03
Visual axis opacification	42 (0.84)		1 (0.04)		0.80 (0.64, 0.96)	<0.001
Additional laser surgery	42 (0.84)		0		0.84 (0.71, 0.97)	<0.001
Anterior vitrectomy	8 (0.16)		0		0.16 (0.03, 0.29)	0.04

Summary statistics of decimal acuity measured at each time point and complication in infants who received the new surgical technique versus the standard of care. Mean (standard deviation) is reported for continuous variables and count (percentage) is reported for categorical variables. CI, confidence interval; logMAR, logarithm of the minimum angle of resolution. OD, *oculus dexter* (right eye); OS, *oculus sinister* (left eye).

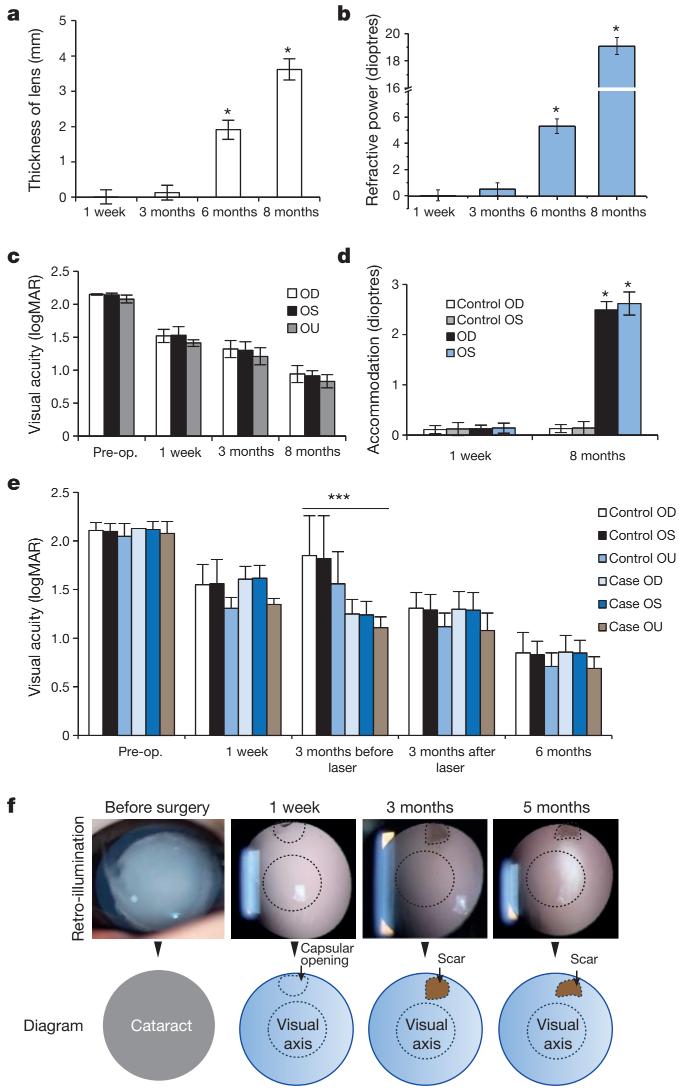


Figure 5 | Functional characteristics of regenerated human lenses. **a**, Lens thickness increased significantly 6 and 8 months after surgery (1.9 ± 0.3 and 3.7 ± 0.3 mm, respectively, $*P < 0.01$), $n = 24$. **b**, Lens refractive power increased significantly 6 and 8 months after surgery (5.1 ± 0.5 and 19.0 ± 0.6 dioptres, respectively, $*P < 0.01$). $n = 24$. **c**, Visual acuity improved after surgery. logMAR, logarithm of the minimum angle of resolution. Pairwise analysis was performed to compare visual acuity before and after surgery ($P < 0.05$) OD (*oculus dexter*, right eye) $n = 25$; OS (*oculus sinister*, left eye) $n = 12$, OU (*oculus uterque*, both eyes) $n = 12$. **d**, Accommodative power increased significantly from 1 week (control OD, control OS, OD and OS, all 0.1 ± 0.1 dioptres) to 8 months (control OD and control OS, 0.2 ± 0.1 dioptres; OD and OS, 2.5 ± 0.2 dioptres) postoperatively ($*P < 0.001$). Control OD, $n = 25$; control OS, $n = 12$; OD, $n = 12$; OS, $n = 12$. **e**, Visual acuity was measured preoperatively and at 1 week, 3 months, and 6 months postoperatively. The majority of eyes in the control group underwent additional laser capsulotomy at 3 months after surgery, with visual acuity measured before and after the procedure. There was no significant difference in visual acuity between eyes that received minimally invasive surgery ($n = 24$) and those that received the current surgical technique ($n = 50$), except at 3 months before the control group underwent laser capsulotomy (t-test, $***P < 0.001$). Data are shown as mean ± s.d. **f**, Visual axis transparency was achieved in nearly all cataractous infant eyes after minimally invasive surgery (95.8%). The scar tissue of the wound on the anterior capsule was <1.5 mm in diameter and located in the periphery, away from the visual axis. The scars were not visible unless the pupils were dilated. No disorganized tissue regeneration was observed. Compared with the current standard surgical method, the new surgical technique decreased VAO by >20-fold.

regenerated lenses, which was markedly improved compared to the 0.10 dioptre increase in aphakic controls ($P < 0.001$). Grating acuities (cycles/degree) was recorded preoperatively and at each postoperative follow-up appointment, then converted to the logarithm of the minimum angle of resolution. Infant visual acuity and accommodation power were significantly improved postoperatively compared to the preoperative baseline ($P < 0.001$) (Fig. 5c, d). The increase in visual acuity was comparable to that achieved using the current surgical method (Fig. 5e). Thus, visual function testing showed that the regenerated lenses were functional.

Clinical outcome with new or current treatment approach

It is well known that with the current method for paediatric cataract surgery, VAO will occur in nearly all patients weeks or months postoperatively owing to the abnormal proliferation of residual LECs[20] (Table 1). To avoid VAO, additional procedures such as polishing of the lens capsule, laser capsulotomy, PCCC, and anterior vitrectomy are widely practiced to disrupt LECs, the lens capsule on which LECs proliferate, and aberrant lens fibre regeneration. Although these procedures can decrease VAO incidence by 15%, they carry significant risk of postoperative inflammation and complications. In this clinical trial, our minimally invasive surgical method resulted in visual axis transparency in nearly all eyes (95.8%) (Fig. 5f, Table 1). Since the scar from the ACCC was <1.5 mm in diameter and located in the periphery of the anterior capsule, it was far from the visual axis (Fig. 5f) and not visible unless the pupils were dilated. The preserved lens capsule remained nearly entirely transparent (Fig. 5f). No disorganized tissue regeneration was observed. Thus, compared to the current standard of care for cataract surgery, our new minimally invasive technique decreased VAO by more than 20-fold (84% versus 4.2%). Furthermore, there was an intact posterior capsule and lens–vitreous interface (Table 1).

By using paired t-tests within each group, significant improvement of decimal acuity before and after treatment was observed with $P < 0.001$ ($t = 23.40$, degrees of freedom (d.f.) = 49.04) in the standard-of-care group and $P < 0.001$ ($t = 15.05$, d.f. = 23.01) in the novel treatment group. A linear mixed-effect model using decimal acuity as the outcome (time: baseline, 1 week, 3 month (after surgery for control group)) and treatment assignment and their interaction as fixed effects yielded statistically insignificant result for time and treatment interaction by likelihood ratio test with $P = 0.956$ ($\chi^2 = 0.332$, d.f. = 3) (Extended Data Table 1a, left, and Extended Data Table 1c, left)), which indicated that the mean response profiles for the two groups were parallel over time. We then refit the linear mixed-effect model by dropping out the interaction term (Extended Data Table 1b, left). A likelihood ratio test with insignificant P value of 0.776 ($\chi^2 = 0.081$, d.f. = 1) (Extended Data Table 1c, left), suggested that the difference between mean decimal acuity in two groups was not statistically different over time (Extended Data Fig. 8b). In contrast, the linear mixed-effect model using decimal acuity as the outcome (time: baseline, 1 week, 3 month (before surgery for control group)) and treatment assignment and their interaction as fixed effects yielded statistically significant results for time and treatment interaction with $P < 0.001$ ($\chi^2 = 47.529$, d.f. = 3) (Extended Data Table 1a, right, and Extended Data Table 1c, right). The non-parallel pattern of mean responses from two groups was largely due to vision loss at 3 months before laser surgery in the control group, while the decimal acuity was monotonically increased in the novel treatment group (Extended Data Fig. 8b). The novel treatment also shows significantly lower complication rate by almost every measurement, supporting the superiority and safety of the novel treatment (Table 1).

Discussion

Each year, more than 20 million cataract patients worldwide undergo treatment with lens extraction and artificial IOL implantation[5]. Despite the clinical success of IOLs, they have numerous limitations and potential complications, including IOL dislocation, suboptimal biocompatibility, inadequate accommodation, and poor visual outcomes. In the

worst cases, irreversible blindness may result[21]. Thus, a new strategy for treating congenital cataracts using naturally regenerated lenses is highly desirable.

The current surgical procedure for paediatric cataracts impairs lens regeneration in several ways. First, the commonly used ACCC creates a relatively large opening at the centre of the anterior capsule, prolonging recovery time and increasing the incidence of inflammation, while wound healing may form scars and cause postoperative VAO. Second, the surgical procedure removes most of the anterior subcapsular LECs, of which a subpopulation is critical for lens regeneration[22]. Third, abnormal proliferation of residual LECs causes postoperative VAO in many cases[23], which requires opening of the posterior capsule, performed by either laser capsulotomy or PCCC and anterior vitrectomy[24]. By destroying the integrity of the lens capsule and LECs, the current surgical procedure greatly diminishes the possibility of lens regeneration[25].

In this study, we show that a new minimally invasive cataract surgery method preserves the integrity of the lens capsule and associated LECs, facilitating functional lens regeneration in animals and humans. In addition to achieving lens regeneration in patients with congenital cataracts, our method also increased visual axis transparency and decreased the rate of complications. The small capsulorhexis opening healed quickly with a nearly intact lens capsule and minimal postoperative inflammation (Fig. 5f, Extended Data Fig. 8c–e). After surgery, there was a clear cornea, anterior chamber, and fundus, with no surgery-related complications.

Our method resulted in visual axis transparency in >95% of cataractous eyes in infants, a much higher percentage than that obtained by traditional surgery. In the remaining patient with VAO, some degree of opacification of the regenerated lens occurred that correlated with improper healing of the anterior capsulorhexis opening and loss of LECs. Therefore, visual axis transparency necessitates preservation of the integrity of the lens capsule and associated LECs, as was illustrated by our lens regeneration study in rabbits. Furthermore, we found that BMI-1 is required for maintenance and renewal of endogenous LECs, and that loss of BMI-1 leads to reduction of the proliferative capacity of LECs and cataract formation. The fact that LECs in adult human eyes exhibit increased proliferative potential after injury highlights the potential for LEC replenishment beyond the paediatric population. These findings may therefore have implications for lens regeneration in elderly patients with age-related cataracts. However, there are important differences between paediatric and adult cataracts. Hard cataracts (nuclear sclerosis) in adults may require phacoemulsification, which could damage LECs; in addition, tissue consistency and capsular thickness/elasticity may pose other challenges for adult lens regeneration. Furthermore, differences between the regenerative capacities of paediatric and adult lenses may suggest a prolonged period of regeneration in adults.

In summary, the current surgical procedure for cataract treatment inadvertently destroys the integrity of the lens capsule and the very LECs that hold the regenerative key to lens restoration. It is also associated with numerous side-effects and a significant risk of complications, particularly in infants. To overcome these problems, we have developed a new, minimally invasive surgical method that allows regeneration of a functional lens with refractive and accommodative abilities, and with greater visual axis transparency. This new cataract treatment uses endogenous stem cells to replenish the human ocular lens, and provides a fresh paradigm for organ and tissue regeneration.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 October 2014; accepted 29 January 2016.
Published online 9 March 2016.

1. Barbosa-Sabanero, K. *et al.* Lens and retina regeneration: new perspectives from model organisms. *Biochem. J.* **447,** 321–334 (2012).
2. Tsonis, P. A. & Del Rio-Tsonis, K. Lens and retina regeneration: transdifferentiation, stem cells and clinical applications. *Exp. Eye Res.* **78,** 161–172 (2004).
3. Gwon, A. Lens regeneration in mammals: a review. *Surv. Ophthalmol.* **51,** 51–62 (2006).
4. Gwon, A. E., Gruber, L. J. & Mundwiler, K. E. A histologic study of lens regeneration in aphakic rabbits. *Invest. Ophthalmol. Vis. Sci.* **31,** 540–547 (1990).
5. Stevens, G. A. *et al.* Global prevalence of vision impairment and blindness: magnitude and temporal trends, 1990–2010. *Ophthalmology* **120,** 2377–2384 (2013).
6. Lois, N., Taylor, J., McKinnon, A. D. & Forrester, J. V. Posterior capsule opacification in mice. *Arch. Ophthalmol.* **123,** 71–77 (2005).
7. Visser, N., Bauer, N. J. & Nuijts, R. M. Toric intraocular lenses: historical overview, patient selection, IOL calculation, surgical techniques, clinical outcomes, and complications. *J. Cataract Refract. Surg.* **39,** 624–637 (2013).
8. Mamalis, N., Davis, B., Nilson, C. D., Hickman, M. S. & Leboyer, R. M. Complications of foldable intraocular lenses requiring explantation or secondary intervention–2003 survey update. *J. Cataract Refract. Surg.* **30,** 2209–2218 (2004).
9. Jacobi, P. C., Dietlein, T. S. & Konen, W. Multifocal intraocular lens implantation in pediatric cataract surgery. *Ophthalmology* **108,** 1375–1380 (2001).
10. Bothun, E. D. *et al.* One-year strabismus outcomes in the Infant Aphakia Treatment Study. *Ophthalmology* **120,** 1227–1231 (2013).
11. Infant Aphakia Treatment Study Group. A randomized clinical trial comparing contact lens with intraocular lens correction of monocular aphakia during infancy: grating acuity and adverse events at age 1 year. *Arch. Ophthalmol.* **128,** 810–818 (2010).
12. Beebe, D. C., Holekamp, N. M. & Shui, Y. B. Oxidative damage and the prevention of age-related cataracts. *Ophthalmic Res.* **44,** 155–165 (2010).
13. Park, I. K. *et al.* Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells. *Nature* **423,** 302–305 (2003).
14. Lessard, J. & Sauvageau, G. *Bmi-1* determines the proliferative capacity of normal and leukaemic stem cells. *Nature* **423,** 255–260 (2003).
15. Molofsky, A. V. *et al. Bmi-1* dependence distinguishes neural stem cell self-renewal from progenitor proliferation. *Nature* **425,** 962–967 (2003).
16. Tsonis, P. A. *Animal models in eye research.* 1st edn, (Academic Press, 2008).
17. Gogate, P., Kalua, K. & Courtright, P. Blindness in childhood in developing countries: time for a reassessment? *PLoS Med.* **6,** e1000177 (2009).
18. Wilson, M. E., Saunders, R. A. & Trivedi, R. H. *Pediatric Ophthalmology: Current Thought and A Practical Guide.* (Springer-Verlag, Berlin, 2009).
19. You, C. *et al.* Visual impairment and delay in presentation for surgery in chinese pediatric patients with cataract. *Ophthalmology* **118,** 17–23 (2011).
20. Wilson, M. E., Trivedi, R. H. & Pandey, S. K. *Pediatric Cataract Surgery: Techniques, Complications, and Management.* (Lippincott Williams & Wilkins, 2005).
21. Zheng, Q. *et al.* Vitreous surgery for macular hole-related retinal detachment after phacoemulsification cataract extraction: 10-year retrospective review. *Eye (Lond.)* **26,** 1058–1064 (2012).
22. Zhou, M., Leiberman, J., Xu, J. & Lavker, R. M. A hierarchy of proliferative cells exists in mouse lens epithelium: implications for lens maintenance. *Invest. Ophthalmol. Vis. Sci.* **47,** 2997–3003 (2006).
23. Plager, D. A. *et al.* Complications, adverse events, and additional intraocular surgery 1 year after cataract surgery in the Infant Aphakia Treatment Study. *Ophthalmology* **118,** 2330–2334 (2011).
24. Nihalani, B. R. & VanderVeen, D. K. Technological advances in pediatric cataract surgery. *Semin. Ophthalmol.* **25,** 271–274 (2010).
25. Gwon, A., Gruber, L. J. & Mantras, C. Restoring lens capsule integrity enhances lens regeneration in New Zealand albino rabbits and cats. *J. Cataract Refract. Surg.* **19,** 735–746 (1993).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Yeh, L. Xi, J. Shelton, A. Pineda and R. Ufret-Vincenty for technical assistance. This study was funded by 973 Program (2015CB964600, 2014CB964900, 2013CB967504); a Major International Joint Research Project (No. 81320108008); 863 Program (2014AA021604), NSFC (No. 81270981); the State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University; Research to Prevent Blindness and the Howard Hughes Medical Institute.

Author Contributions H.L., S.H., Z.L., S.C., .X.L., L.L., B.C., Y.W. and Y.L. conducted the clinical trial; H.O., Jie Z., Y.Z., J.C., H.C. and S.P. performed mouse LEC lineage tracing experiments; H.O., J.Z., G.C., G.L., Y.X., S.P., Jin Z., M.J., A.Q., F.S., X.L., R.H., W.S. and D.C. performed LEC characterization and differentiation experiments; D.G., C.H., F.W., Z.S. and J.H. analysed clinical trial data; H.O., M.K., D.S-K., C.C., M.J., Y.W., W.S., D.C., S.Z., L.Z. and S.C. performed gene expression studies and analysed data; R.A.J.S. and S.J.M. performed and analysed the experiments related to BMI-1 function in mouse lens epithelium. Y.L, R.M. and K.Z. designed the study and wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.L. (yzliu62@yahoo.com) or K.Z. (kang.zhang@gmail.com).

METHODS

Isolation and culture of LECs. All animal studies were performed with the approval of the Institutional Animal Care Committees of Sun Yat-sen University, the University of California San Diego, West China Hospital, and the University of Texas Southwestern Medical Center.

The eyeball was enucleated from a one-month-old New Zealand white rabbit and washed with PBS (containing antibiotics) three times. After the cornea and iris were removed, a small cut was made in the posterior capsule of the lens; the capsule with attached epithelium was removed and cut into $1 \times 1\,\text{mm}^2$ pieces. The pieces of epithelium were cultured in minimum essential media supplemented with 20% FBS, NEAA, and $50\,\mu\text{g}\,\text{ml}^{-1}$ gentamicin.

A 17-week-old human fetal eyeball was purchased from Advanced Bioscience Resources, Inc. (San Francisco, California). Post-mortem human eyes were obtained from San Diego Eye Bank. The human LECs were cultured according to the same methods as above.

For *in vitro* differentiation, LECs were cultured on Matrigel-coated six-well plates or eight-well chambers. Lentoid body was formed after 21 days in minimum essential media supplemented with NEAA, 1% FBS, $100\,\text{ng}\,\text{ml}^{-1}$ FGF2, and $5\,\mu\text{g}\,\text{ml}^{-1}$ insulin. Images of lentoid tissue were obtained using a Leica M205FA stereo microscope.

Transgenic mouse study. Membrane-tomato/membrane-green (mTmG)-targeted $ROSA^{mTmG}$ mice were purchased from the Jackson Laboratory (Bar Harbour, ME; stock no. 7576) and maintained as homozygotes. *P0-3.9-GFPcre* mice expressing an eGFP–Cre recombinase fusion protein under the control of the *Pax6* lens ectoderm enhancer and the *Pax6 P0* promoter[26] were maintained in a FVB/N background. Lineage-tracing experiments were performed by crossing the homozygous $ROSA^{mTmG}$ reporter mouse strain with the *P0-3.9-GFPcre* deleter strain. Eyes were dissected at P1, P14, and P30 and fixed overnight in 4% formaldehyde. Tissues were then incubated in 10% sucrose and embedded in OCT for cryo-sectioning. Frozen sections were washed in PBS and imaged on a Zeiss Axio Imager fluorescence microscope. $Bmi1^{fl/fl}$ mice were generated as previously described[27]. *Nestin-cre* mice[28] were obtained from the Jackson Laboratory. For BrdU pulses, mice were injected with $100\,\text{mg}\,\text{kg}^{-1}$ BrdU (Sigma) dissolved in PBS, then maintained on drinking water that contained $1\,\text{mg}\,\text{ml}^{-1}$ BrdU until sacrifice.

For gene expression studies, lenses of *Pax6P0-3.9-GFPcre* mice were dissected under a dissecting microscope. Lens capsular bag was opened from the posterior surface by making three crisscross incisions. The capsular bag was opens and lens material extruded. GFP-positive LECs in the mid-anterior capsular area were separated mechanically from GFP-negative LECs in the remaining capsular areas under a fluorescence microscope. RNA was isolated using RNeasy Mini Kit (Qiagen).

To image cataracts, mice were anaesthetized with Avertin, and one drop of 1% Mydriacyl (Alcon) was administered per eye. Eyes were immediately visualized *in vivo* using a light microscope. For histology, mice were perfused with heparinized saline followed by 4% paraformaldehyde (PFA) in PBS. Dissected eyes were fixed in 4% PFA overnight, embedded in paraffin, and sectioned by the UT Southwestern Molecular Pathology core facility. For BrdU staining, slides were deparaffinized, and subjected to heat-mediated antigen retrieval (in 10 mM sodium citrate, pH 6.0). Slides were stained with primary mouse anti-BrdU (Caltag, MD5000, 1:200) overnight at 4 °C. Slides were subsequently stained with Alexa Fluor 555-conjugated goat anti-mouse IgG1 secondary antibody (Life Technologies, 1:500) and $1\,\text{mg}\,\text{ml}^{-1}$ DAPI (1:500) for 1 h at room temperature. The number of BrdU-labelled cells was divided by the total number of DAPI^{+} cells in a single layer of LECs.

Lentiviral RNAi. Lentiviral shRNA targeting the human *BMI1* gene (NCBI Reference Sequence: NM_005180.8) was purchased from Origene (TL314462), ShRNA targeting sequences were as follow: 5′-AATGCCATATTGGTATATGACATAACAGG-3′ and 5′-GTAAGAATCAGATGGCATTATGCTTGTTG-3′. Two shRNAs were used separately, and a non-effective 29-mer scrambled shRNA was used as a control. Lentiviral shRNA particles were prepared using shRNA lentiviral packaging kit (Origene, TR30022). Viruses were harvested at 48 h and 72 h post-transfection.

Western blot analysis. LECs were cultured on Matrigel-coated 3.5-mm dishes with lentoid formation medium for 30 days. Cells were washed twice with ice-cold PBS, and lysed in RIPA lysis buffer with PMSF. Protein concentration was determined by BCA protein assay kit. Thirty micrograms of total protein lysate was loaded onto 10% SDS–PAGE gel and then transferred to a PVDF membrane (Millipore) at 70 V for 2 h. The membrane was probed with the following primary antibody at 4 °C overnight: anti-αA-crystallin (sc-22389, Santa Cruz), anti-β-crystallin (sc-48335, Santa Cruz), anti-γ-crystallin (sc-22415, Santa Cruz) and anti-β-actin (sc-47778, Santa Cruz), and then incubated with HRP-conjugated anti-rabbit, anti-mouse, or anti-goat secondary antibody for 1 h at room temperature. The immunodetection was visualized using a blot imaging system (Fluor Chem Q, Protein Simple) with ECL buffer (Millipore).

Lens regeneration in rabbit and macaque models. New Zealand white rabbits ($n = 29$, four rabbits died from systemic infections unrelated to surgery. The remaining 25 rabbits were used to assess regeneration), and long-tailed macaques (*Macaca fascicularis*) monkeys ($n = 6$) underwent minimally invasive capsulorhexis surgery. Only the left eye of each animal was used for experiments. Slit-lamp biomicroscopy and photography were performed at different time points to monitor lens regeneration. Rabbits were euthanized at day 1, day 7, and one month after surgery, and the treated eyes were enucleated. The lenses were harvested for histologic analysis using haematoxylin and eosin staining. For the macaques, enucleation of the treated eye was performed 4 months post-surgery and the lenses were harvested for the same histologic examinations. The eyes were fixed, paraffin-embedded, and sectioned at $5\,\mu\text{m}$ through the cornea, pupil, and optic nerve with the lens *in situ*.

Real-time PCR. RNA was isolated from rabbit LECs, mature lens fibre cells and LECs in *P0-3.9-GFPcre* mice using an RNeasy Mini Kit (Qiagen) and subjected to on-column DNase digestion. cDNA was synthesized using a Superscript III reverse transcriptase kit according to the manufacturer's instructions (Invitrogen). Quantitative PCR was performed via 40 cycle amplification using gene-specific primers (Supplementary Table 1) and Power SYBR Green PCR Master Mix on a 7500 Real-Time PCR System (Applied Biosystems). Measurements were performed in triplicate and normalized to endogenous GAPDH levels. The relative fold change in expression was calculated using the $\Delta\Delta C_t$ method (C_t values <30).

Immunofluorescence and laser confocal microscopy. Rabbit LECs were fixed in 4% PFA for 20 min, then permeabilized with 0.3% Triton X-100-PBS for 10 min and blocked in PBS solution containing 5% BSA, followed by an overnight incubation in primary antibodies at 4 °C. After three washes in PBS, cells were incubated with secondary antibody for 1 h in room temperature. Cell nuclei were counterstained with DAPI.

The following antibodies were used: goat anti-Sox2 polyclonal antibody (Santa Cruz), rabbit anti-PAX6 polyclonal antibody (PRB-278P, Covance), mouse anti-Bmi1 antibody (ab14389, Abcam), and mouse anti-Ki67 monoclonal antibody (550609, BD Sciences). The secondary antibodies, Alexa Fluor 488- or 568-conjugated anti-mouse or anti-rabbit IgG (Invitrogen), were used at a dilution of 1:500. Images were obtained using an Olympus FV1000 confocal microscope.

BrdU labelling of LECs in humans. We used BrdU labelling to identify and quantify proliferating LECs from human cadaver eyes. Whole-mount human lens capsules were pulsed with BrdU and then stained with an antibody against BrdU to determine the distribution and density of proliferating LECs. In brief, within 12–24 h after death, lenses from post-mortem donor eyes were obtained from the Eye Bank of Zhongshan Ophthalmic Center in Guangzhou, China. Twelve lenses in total from six donors were used for the experiment. A small puncture injury was made on the anterior surface of a post-mortem human lens using a 30-gauge needle. The lenses were cultured at 37 °C in Dulbecco modified Eagle medium (DMEM) supplemented with 10% FBS in a humidified incubator with 5% CO_2. The contralateral lens from the same donor was treated under the same conditions but did not receive a puncture injury and was used as a control. To label the proliferating LECs, both groups of lenses were incubated in $100\,\mu\text{g}\,\text{ml}^{-1}$ BrdU (Sigma-Aldrich) 24 h after the puncture injury. The lens was then removed from the capsular bag, and the lens capsules were fixed in 4% formaldehyde and subjected to BrdU staining using a standard immunohistochemistry protocol according to the manufacturer's instructions (CST, Boston, Massachusetts). Images were taken using a Carl Zeiss microscope (Jena, Germany).

Study design, execution, and oversight of clinical trial in humans. This study was approved by the institutional review board of the Zhongshan Ophthalmic Center (ZOC). Informed written consent was obtained from the parents or guardians of the infants before enrolment, and the tenets of the Declaration of Helsinki were followed throughout the study. The study was conducted in accordance with an international guideline and protocol for visual function measurements in paediatric cataract surgery and a protocol of the Childhood Cataract Program of the Chinese Ministry of Health (CCPMOH) and had an independent data and safety monitoring board of ZOC-CCPMOH.

Description of current surgical method for cataract extraction. The current standard-of-care treatment for paediatric cataract involves removal of the cataractous lens through a relatively large opening using anterior continuous curvilinear capsulorhexis (ACCC, about 6 mm in diameter, Extended Data Fig. 1), followed by cataract extraction and artificial lens implantation or placement of postoperative aphakic eyeglasses/contact lens in paediatric cataract patients younger than two years. Some patients underwent additional posterior continuous curvilinear capsulorhexis (PCCC) and anterior vitrectomy.

Establishment of a minimally invasive capsulorhexis surgery method to preserve LECs. We established a new capsulorhexis surgery method to facilitate lens regeneration (Fig. 3a). First, we decreased the size of the capsulorhexis opening to 1.0–1.5 mm in diameter. This results in a minimal wound of about $1.2\,\text{mm}^2$ in area, which is only about 4.3% the size of the wound created by the current

method. Second, we moved the location of the capsulorhexis to the peripheral area of the lens instead of the central area. A 0.9 mm phacoemulsification probe was used to remove the lens contents and/or cortical opacities. These changes provide significant advantages. First, it considerably reduces the size of the injury, which resulted in a lower incidence of inflammation and much faster healing. Second, it moves the wound scar away from the central visual axis to the periphery, leading to improved visual axis transparency. Third, it preserves a nearly intact transparent lens capsule and layer of LECs, which have regenerative potential and are critically required for the regeneration of a natural lens.

Clinical trial of minimally invasive lens surgery in human infants with congenital cataract. The clinical trial is an open label, randomized controlled trial in a study population of paediatric cataract patients (age: 0–2 years). Except the trial participants, all other parties (care providers, outcome assessors) were blinded to treatment allocation. A clinical trial consort flowchart is listed in the Extended Data Fig. 8a. Paediatric patients were enrolled accordingly inclusion and exclusion criteria below (ClinicalTrials.gov identifier: NCT01844258). Inclusion criteria were the following: infants were $\leq$24 months old, and diagnosed with bilateral uncomplicated congenital cataract with an intact non-fibrotic capsular bag. Exclusion criteria included preoperative intraocular pressure (IOP) $>$21 mm Hg, premature birth, family history of ocular disease, ocular trauma, or other abnormalities, such as microcornea, persistent hyperplastic primary vitreous, rubella, or Lowe syndrome. In total, twelve paediatric cataract patients (24 eyes) received the new minimally invasive lens surgery (Table 1). Twenty-five paediatric cataract patients (50 eyes in total) were enrolled as the control group to receive the current standard surgical treatment (Extended Data Fig. 8a). Bilateral eye surgeries of the same patient were conducted during the same operation session.

We defined the incidence of corneal oedema as a $>$5% increase in central corneal thickness one week post-surgery, and the incidence of severe anterior chamber inflammation as a Flare value $>$10 evaluated by Pentacam system (OCULUS, Germany) and Laser flare meter (KOWA FM-600, Japan). Early-onset ocular hypertension was identified as IOP $>$21 mm Hg by Tonopen (Reichert, Seefeld, Germany) within one month after surgery. Macular oedema was identified by fundus OCT (iVue, Optovue, Germany) as an increase in central macular thickness $>$10% one week post-surgery. When indicated, VAO, defined by visual decline and the degree to which the fundus was obscured, was treated with YAG laser capsulotomy at follow-up.

Compared to infants operated on using our new surgical technique, infants who received the traditional technique had a higher incidence of anterior chamber inflammation one week after surgery, early-onset ocular hypertension, and increased VAO (Table 1). However, in the group treated with our new method, a transparent regenerated biconvex lens was found in 100% of eyes three months after surgery, while no regenerated biconvex lenses formed in the group treated with the standard technique. In addition, 100% of the capsular openings healed within one month after surgery in the experimental group, but no capsular openings healed in the control group.

Evaluation of paediatric visual acuity. Testing equipment included a set of Teller Acuity Cards (Vistech Consultants, Dayton, Ohio). The set of cards consists of 15 cards with gratings ranging in spatial frequency from 0.32 to 38 cycles per cm, in half-octave steps, and one blank grey card. A 4-mm peephole in each card allows the tester to view the child's face through the card during testing. Test distance was kept constant by use of an aid to measure the distance from the child's eyes to the card throughout testing. For 38 cm, the aid was the distance measured from the tester's elbow to a specific knuckle on the tester's hand, and for 55 cm, the aid was the length (55 cm) of the Teller Acuity Card. Testers were instructed to hold the cards without wrapping their fingers around the front side of the card, as this may attract the child's attention. Testers presented the cards directly in front of the child and observed the child either over the top of the card or through the peephole in the card.

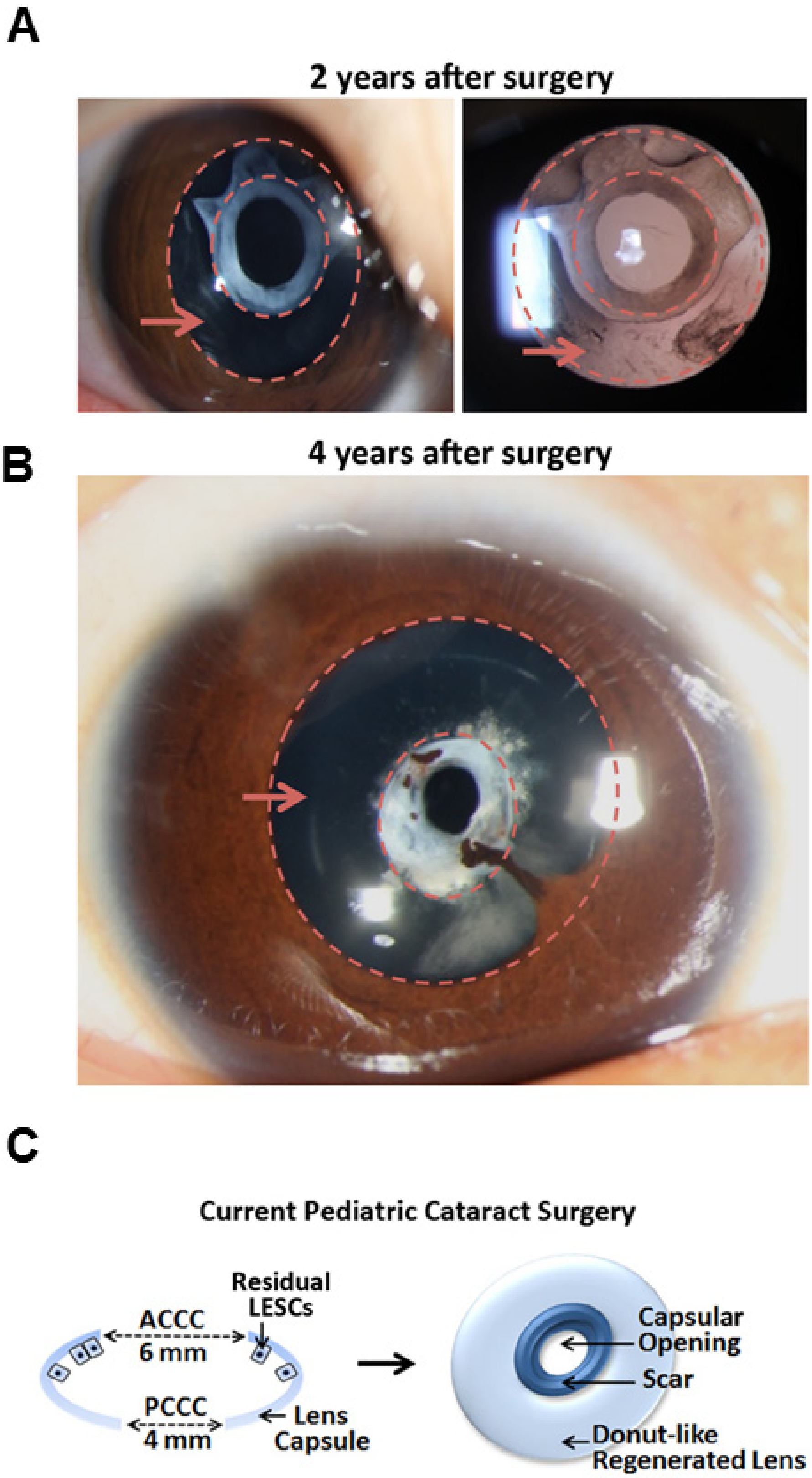
During each acuity test, a masked visual acuity examiner was aware that the gratings were arranged in order from lower to higher spatial frequencies in half-octave steps, but were masked to the absolute spatial frequency of the grating on each card. The subset of spatial frequencies used for each test was selected according to a pseudorandom order from among three possible subsets of spatial frequencies for the subject's age group. All three subsets for each age group included spatial frequencies known to be well above the threshold for that age group. To keep the visual acuity examiner masked to the absolute spatial frequency, the visual acuity examiner was not permitted to look at the front of the card to confirm the location of the grating. Instead, the visual acuity examiner asked an assistant to confirm the location of the grating on the card, after the visual acuity examiner had shown a card to the subject enough times to assess whether or not the subject could detect the grating. A clinical examiner was masked to the acuity results and the assigned patient group. Acuity was scored as the spatial frequency of the finest grating and was converted to log values before data analysis.

Measurement of lens refractive power. We used a handheld auto-refractometer (PlusoptiX A09, OptiMed, Sydney, Australia) to evaluate the function of the regenerated lenses according to the manufacturer's methods.

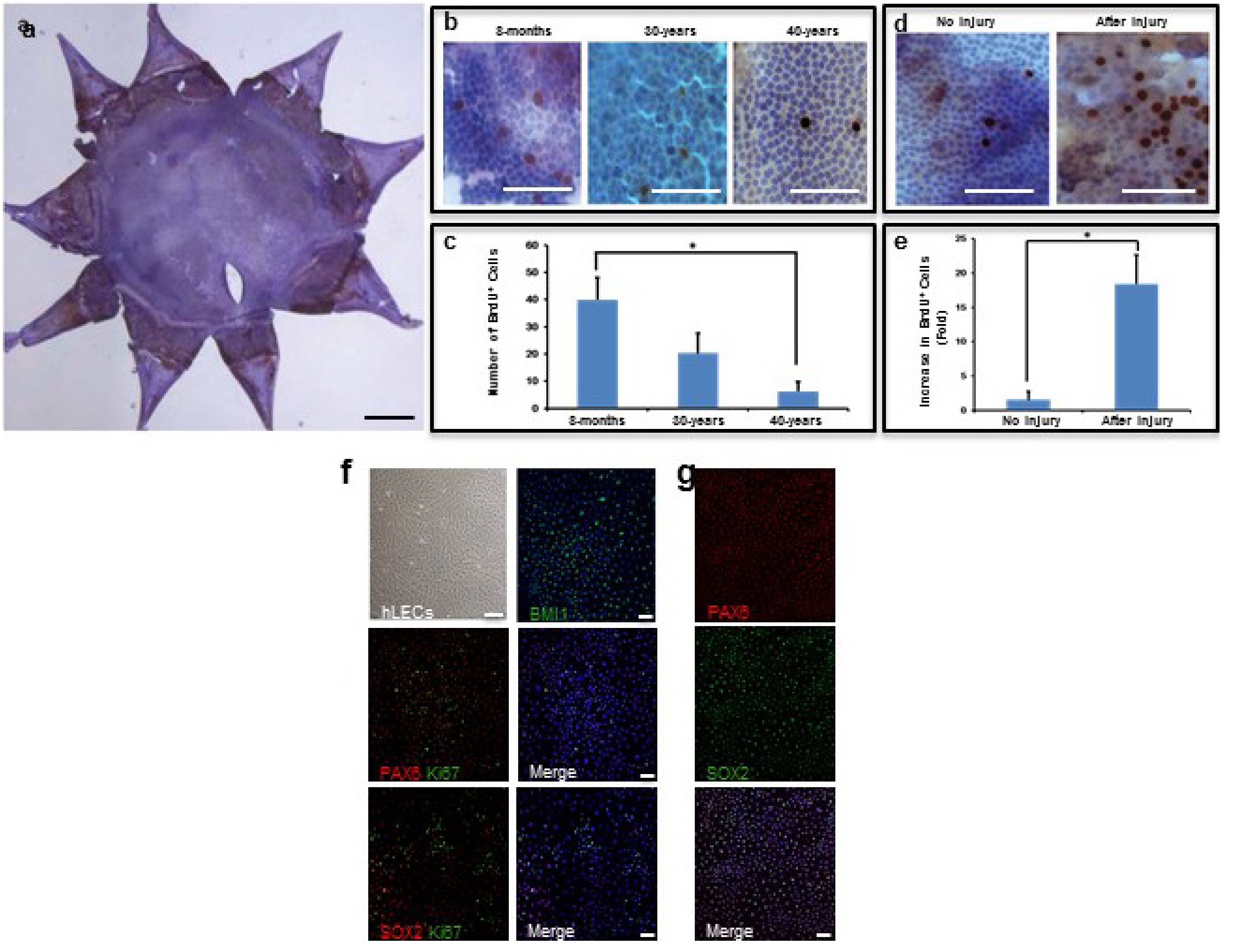
Statistical analysis. Descriptive statistics was provided for the primary and secondary endpoints measured by intervention groups at each time point. Mean and standard deviation was reported for continuous variables and count and percentage is reported for categorical variables. To assess whether the primary outcome, decimal acuity, was significantly improved within each group, we performed the pre-post comparison between decimal acuity measured at baseline and study endpoint using paired t-tests. Normality of the data was checked and non-parametric alternatives, Wilcoxon signed-rank test is considered if the assumption was severely violated. To evaluate whether the mean response profiles in two groups were similar, we used the linear mixed-effect model taking account for the within-subjects correlation. The baseline decimal acuity was not adjusted by the model due to the homogeneity of this measurement as shown in the summary statistics. As the standard-of-care approach requires laser surgery at 3 months while the novel treatment does not, we fit two models using before and after laser surgery data, separately, to demonstrate the superiority of the novel approach. In each model, the outcome is the decimal acuity measured at four time points: baseline, 1 week, 3 months (before or after laser surgery) and 6 months; time (baseline as the reference level), treatment assignment and their interaction are the fixed effects; and patient is the random effect. Significant associations are identified using likelihood ratio test (LRT) by comparing models with and without a fixed effect. A linear mixed-effect model is fit again by dropping out the insignificant fixed effect until the final model is selected. A contrast test is performed when necessary. For the secondary aim, we compared the proportions of each condition of complications between two groups. We assumed the occurrence of complications for eyes from the same patient were independent. The mean difference and its 95% confidence interval was reported. A two-proportion z-test was used with the nonparametric χ^2 test as alternative if the normality assumption was violated. All tests were two-sided and a P value less than 0.05 is considered to be statistically significant.

Evaluation of accommodative response. Accommodative response was measured by an open-field autorefractor (SRW-5001K; Shin-Nippon, Tokyo, Japan), which allows targets to be viewed at any distance. The paediatric patients were positioned for autorefractor measurement with assistance from their parents. The patients were guided to fixate binocularly at a near target (33 cm, 5×5 array of smiley faces of N10 size) and a far target (3 m, 5×5 array of smiley faces of N10 size) by a trained and certified investigator or study coordinator. The measurements from non-cycloplegic autorefraction were performed three times at each target distance by the same trained and certified investigator throughout the study, in order to maintain accuracy and consistency throughout the trial. Measurements were taken in the same quiet environment with consistent room illumination to diminish the influence of distracting factors and to maintain subjects' concentration. The spherical equivalent refractive value (SER) was recorded for each measurement and the mean value was calculated for evaluation of an accommodative response. The value of accommodative response was the difference between SER values for the near and the far target. We also used dynamic retinoscopy to measure the infants' accommodation[29–31]. In brief, we recorded a lens dioptre value using retinoscopy when a patient was guided to fixate on a target 3 m away. Then another lens dioptre value was recorded when the target was moved closer, at a distance of 33 cm from the eyes. The difference between these two measurements was used to evaluate lens accommodative power.

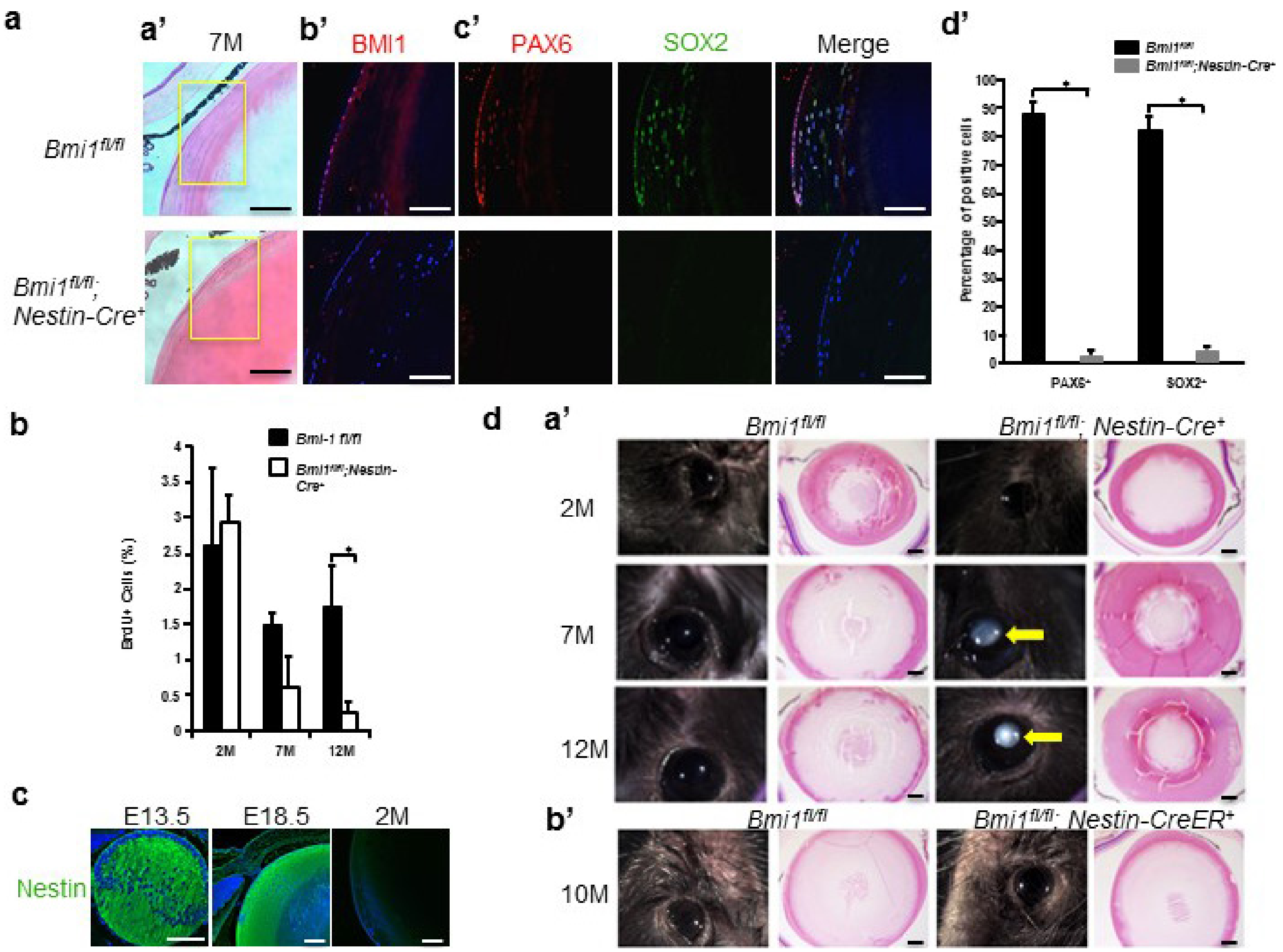
26. Rowan, S. *et al.* Notch signaling regulates growth and differentiation in the mammalian lens. *Dev. Biol.* **321,** 111–122 (2008).
27. Mich, J. K. *et al.* Prospective identification of functionally distinct stem cells and neurosphere-initiating cells in adult mouse forebrain. *Elife* **3,** e02669 (2014).
28. Tronche, F. *et al.* Disruption of the glucocorticoid receptor gene in the nervous system results in reduced anxiety. *Nature Genet.* **23,** 99–103 (1999).
29. Gabriel, G. M. & Mutti, D. O. Evaluation of infant accommodation using retinoscopy and photoretinoscopy. *Optom. Vis. Sci.* **86,** 208–215 (2009).
30. Hainline, L., Riddell, P., Grose-Fifer, J. & Abramov, I. Development of accommodation and convergence in infancy. *Behav. Brain Res.* **49,** 33–50 (1992).
31. Banks, M. S. The development of visual accommodation during early infancy. *Child Dev.* **51,** 646–666 (1980).



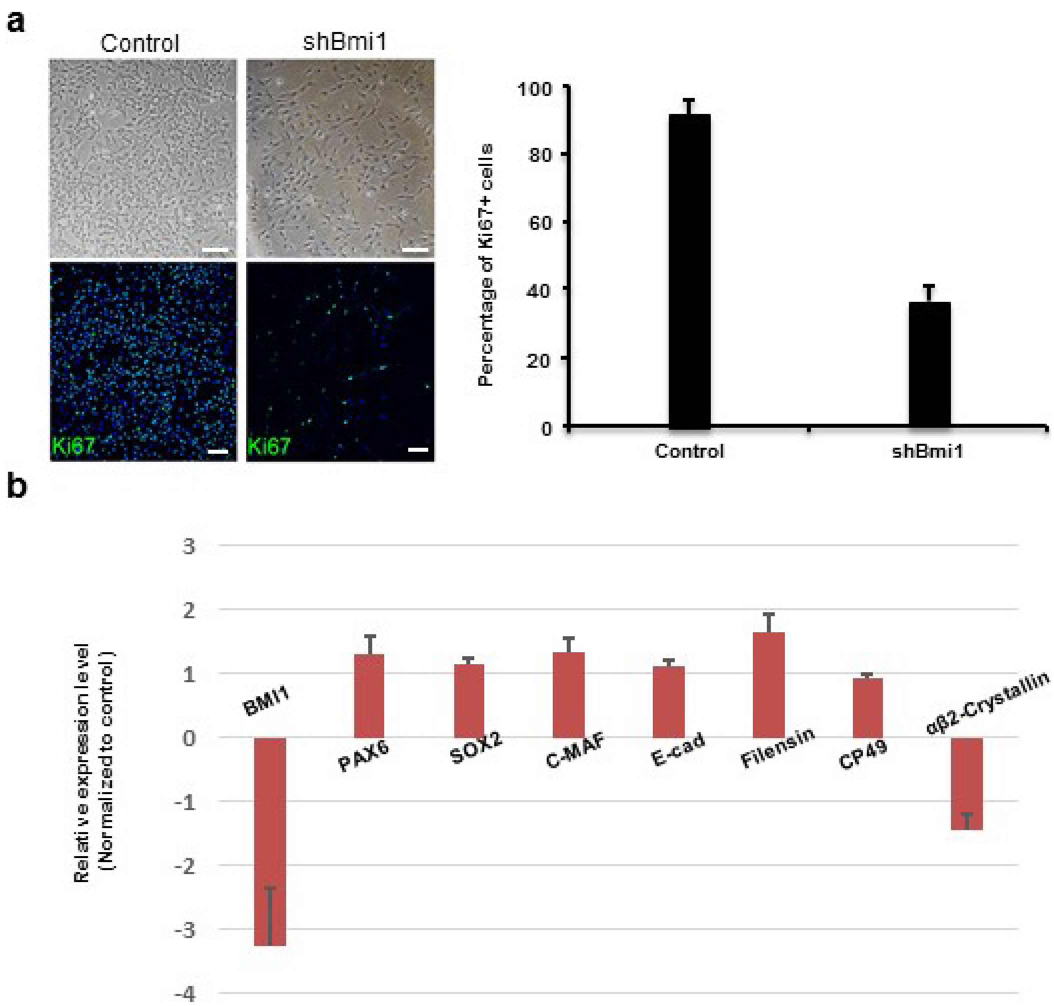
Extended Data Figure 1 | Surgical methods and lens regeneration for congenital cataract. a, **b**, Slit-lamp photography of 'doughnut-like' lens regeneration at different time points after treatment using the current surgical method. Two years after surgery (**a**), the transparent regenerated lens tissue contained the sealed capsular opening with an opaque white scar at the centre. The regions between the dashed circles indicated by the red arrows are the regenerated lens tissues. Four years after surgery (**b**), the capsular opening was constricted compared to that seen at two years post-surgery, indicating continued growth of the regenerated lens. There was also the complication of iridolenticular synechiae. **c**, Schematic diagrams of the current surgical method for paediatric cataracts: the currently practiced paediatric ACCC creates an opening 6 mm in diameter at the centre of the anterior capsule, removing the LECs underneath it and leaving a relatively large wound area of 28 mm^2. The scars formed often cause postoperative VAO. Additionally, PCCC and anterior vitrectomy are commonly performed at follow-up visits.



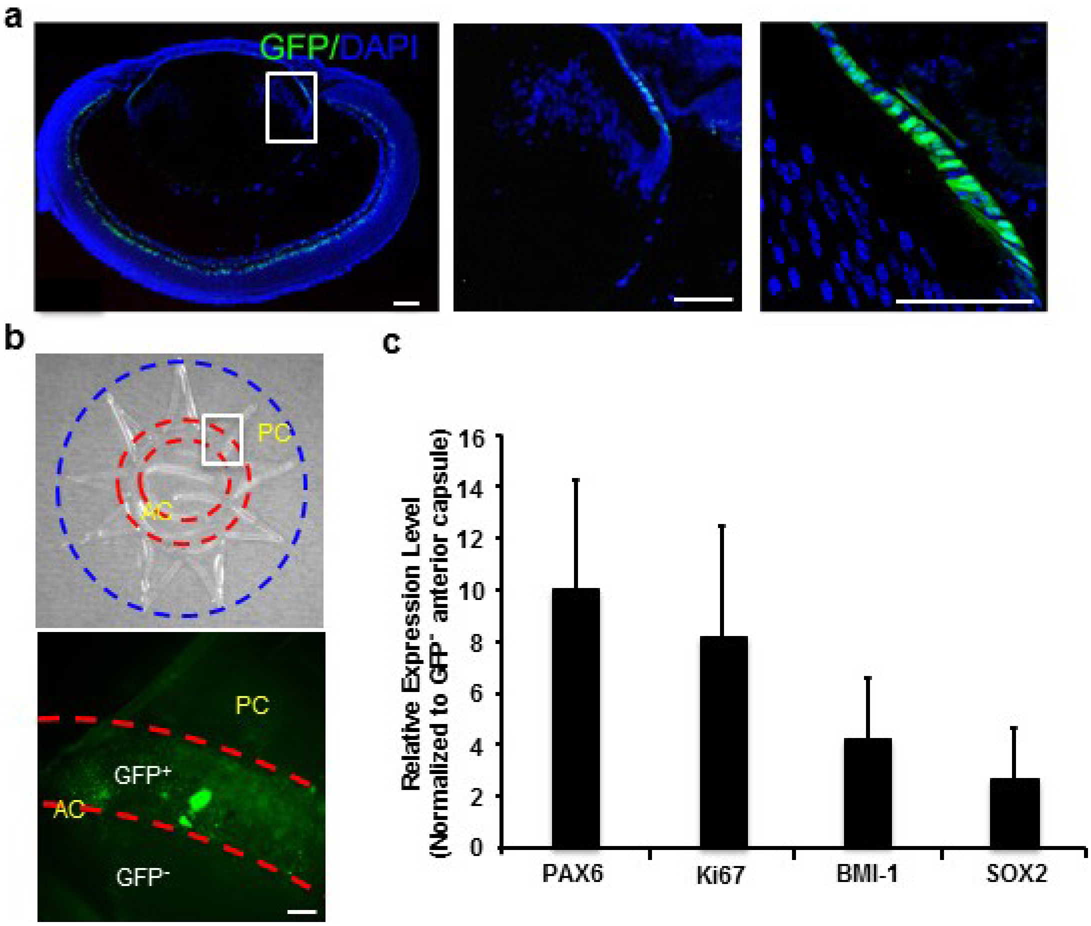
Extended Data Figure 2 | BrdU pulse labelling of human LECs. **a**, Whole mount of a human lens capsule showing BrdU$^+$ cells (brown) by enzymatic immunohistology and diaminobenzidine staining. **b**, High-magnification images of human donor lenses showing BrdU$^+$ LECs. **c**, Bar graph showing quantification of BrdU$^+$ cells. There was an age-dependent decrease in the number of BrdU$^+$ cells (8 months, 39.9 ± 8.1; 30 years, 20.3 ± 7.3 and 40 years, 5.9 ± 2.9; 8 months versus 40 years, $*P < 0.05$). Six randomly chosen fields of each capsule were used for analysis, four samples in each group, ($n = 24$ fields, chosen over four samples). **d**, High-magnification images of whole-mount staining of human lens capsules with or without injury showed a marked increase in the number of BrdU$^+$ cells after injury. **e**, Bar graph showing quantification of BrdU$^+$ cells. The contralateral eyes from the respective donors were used as controls. There was a significant increase in number of BrdU$^+$ cells. No injury, 1.5 ± 1.2; after injury, 18.4 ± 4.2; fold change after injury, 11.3 ± 2.5; $*P < 0.05$. Six randomly chosen fields within the germinative zone of each capsule were used for analysis, five samples in each group ($n = 30$ fields, chosen over five samples). Data shown as means ± s.d. **f**, Cultured human fetal LECs were positive for BMI-1 (green, right upper panel); co-staining of PAX6 (red) and Ki67 (green), middle panels; co-staining of SOX2 (red) and Ki67 (green), lower panels. **g**, Co-staining of PAX6 (red) and SOX2 (green) of human fetal LECs. All scale bars, 100 μm.



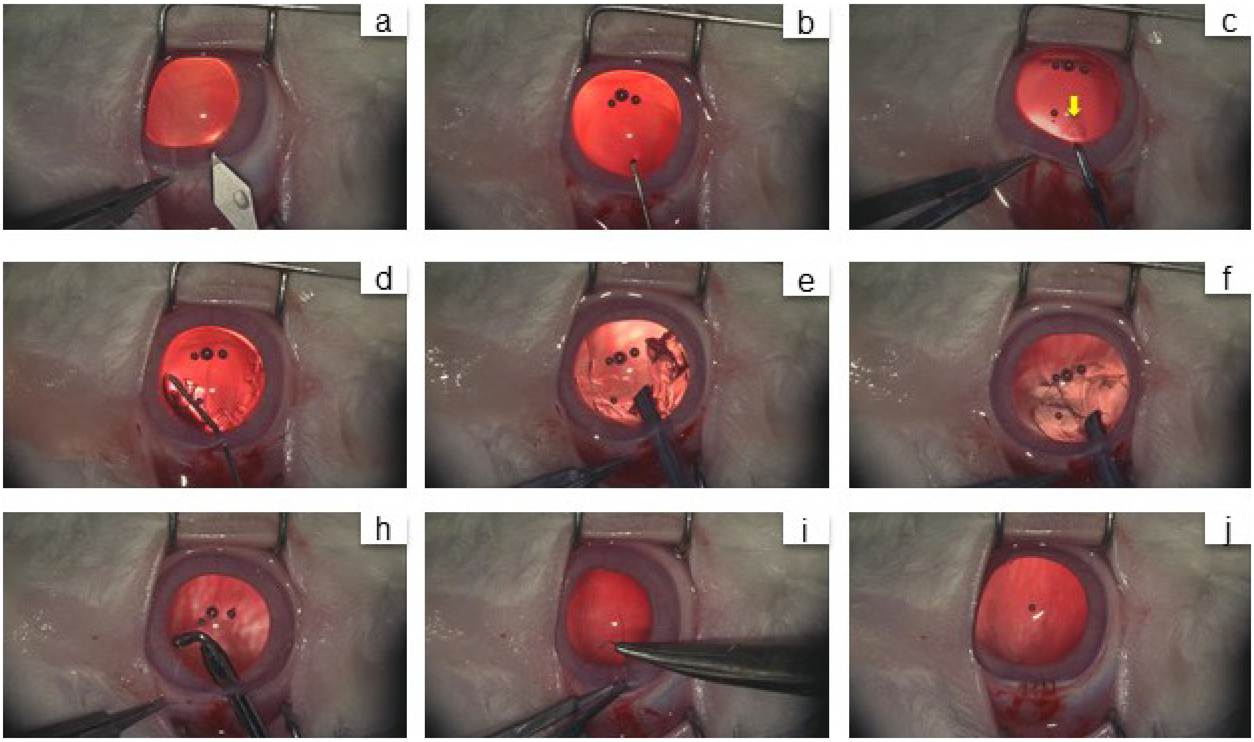
Extended Data Figure 3 | Conditional deletion of Bmi-1 led to decrease in Pax6$^+$ and Sox2$^+$ cells and cataract formation. **A**, Loss of Bmi-1 reduced the Pax6$^+$ and Sox2$^+$ LECs population. **a**, Representative images of haematoxylin and eosin-stained lens sections from *Bmi1$^{fl/fl}$* control mice and *Nestin-cre;Bmi1$^{fl/fl}$* mice. **b**, Representative images of Bmi-1 (red) staining in LECs. **c**, Pax6 (red) and Sox2 (green) immunostaining. **d**, Percentage of Pax6$^+$ (*Bmi1$^{fl/fl}$*, 88.5 ± 2.9%; *Nestin-cre;Bmi1$^{fl/fl}$*, 2.4 ± 2.3%) and Sox2$^+$ (*Bmi1$^{fl/fl}$*, 82.7 ± 3.9%; *Nestin-cre;Bmi1$^{fl/fl}$*, 4.9 ± 1.5%) cells ($n = 5$ mice; 5 sections counted per mice, for a total of 25 sections across 5 mice), $*P < 0.001$. Data are shown as mean ± s.d. **B**, Conditional deletion of *Bmi1* led to reduced LEC proliferation. The percentage of BrdU$^+$ LECs per eye is shown (2M: *Bmi1$^{fl/fl}$*, 2.6 ± 0.9%; *Nestin-cre;Bmi1$^{fl/fl}$*, 3.0 ± 0.4%; $n = 4$ mice. 7M: *Bmi1$^{fl/fl}$*, 1.5 ± 0.2%; *Nestin-cre;Bmi1$^{fl/fl}$*, 0.6 ± 0.4%; $n = 6$ mice. 12M: *Bmi1$^{fl/fl}$*, 1.8 ± 0.6%; *Nestin-cre;Bmi1$^{fl/fl}$*, 0.2 ± 0.2%; $n = 8$ mice), two sections counted per eye. Statistical significance was assessed using a two-tailed Student's *t*-test. $*P < 0.05$. Data are shown as mean ± s.d. **C**, Nestin (green) staining is shown in E13.5, E18.5, and 2-month-old wild-type mice. All scale bars, 100 μm. **D**, Representative images of lenses from *Nestin-cre;Bmi1$^{fl/fl}$* and *Bmi1$^{fl/fl}$* control mice. **a**, Cataracts are evident in 7- and 12-month-old *Nestin-cre;Bmi1$^{fl/fl}$* mice (arrow). **b**, Deletion of Bmi-1 at 6 weeks of age with *Nestn-creER* did not recapitulate the cataract phenotype 10 months after tamoxifen treatment. Haematoxylin and eosin-stained sections of the same eyes are also shown. All scale bars, 100 μm.



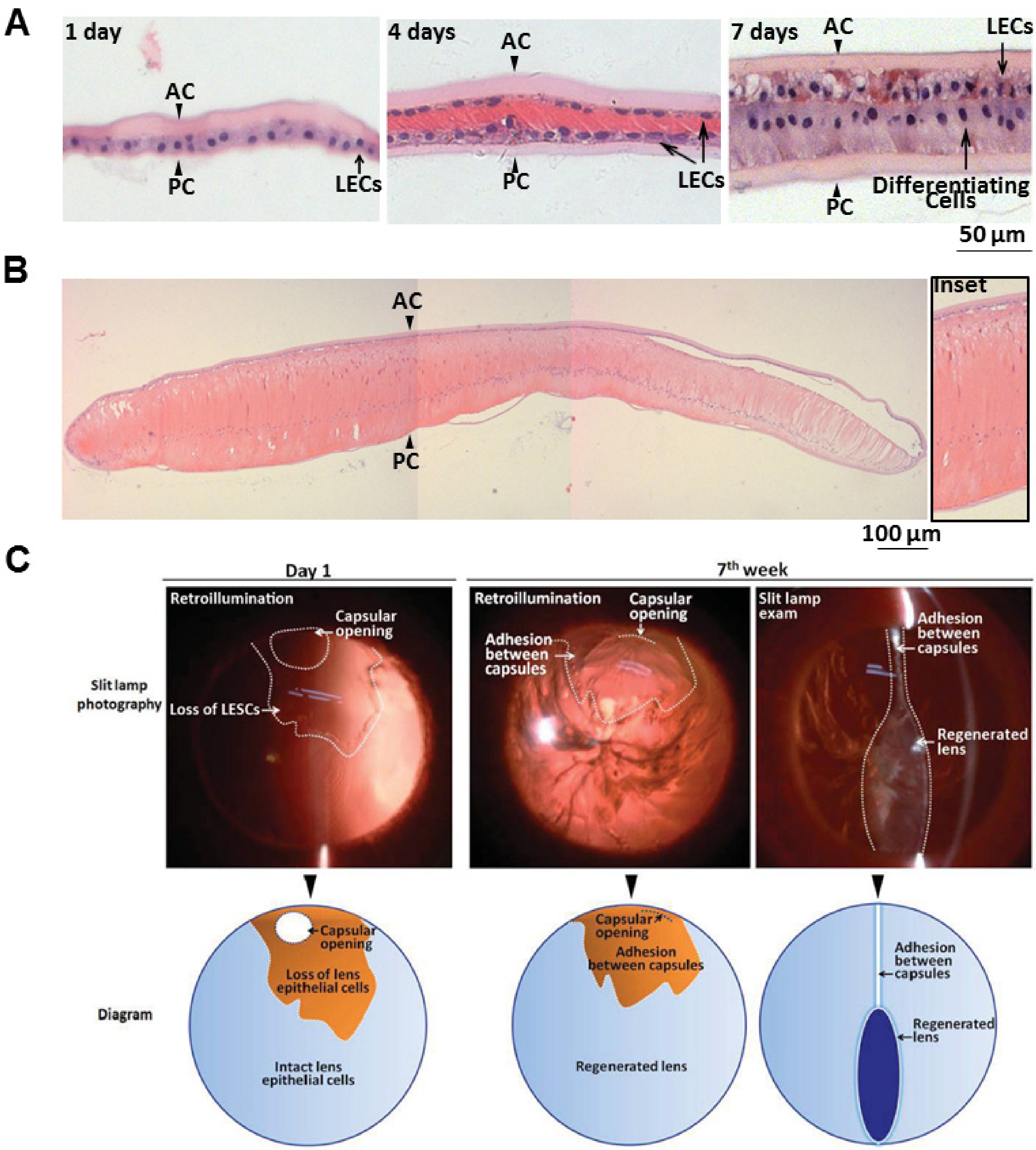
Extended Data Figure 4 | Loss of BMI-1 decreased the proliferative ability of LECs. a, Phase-contrast photographs of human LECs (upper panels) and quantification of Ki67^{+} proliferating human fetal LECs upon *BMI1* knockdown (shBMI1) compared to controls (two shRNAs gave similar results; $n = 5$, $P < 0.05$). Data shown as mean ± s.d. Blue indicates DAPI staining. **b**, Loss of BMI-1 did not significantly affect expression of LEC or lens fibre cell makers in LECs. *BMI1* was reduced by 3.3-fold (all $n = 3$, $P < 0.05$); gene expression changes of LEC markers were: 1.3-fold increase (*PAX6*), 1.1-fold increase (*SOX2*), 1.3-fold increase (*C-MAF*) and 1.1-fold increase (*E-cadherin*); gene expression changes of differentiated lens fibre cell markers were: 1.6-fold increase (*Filensin*), 0.9 fold increase (*CP49*) and 1.4-fold decrease (*CRYBA2*). Two different shRNAs gave similar results; $n = 5$, $P < 0.05$. Data shown as mean ± s.d.



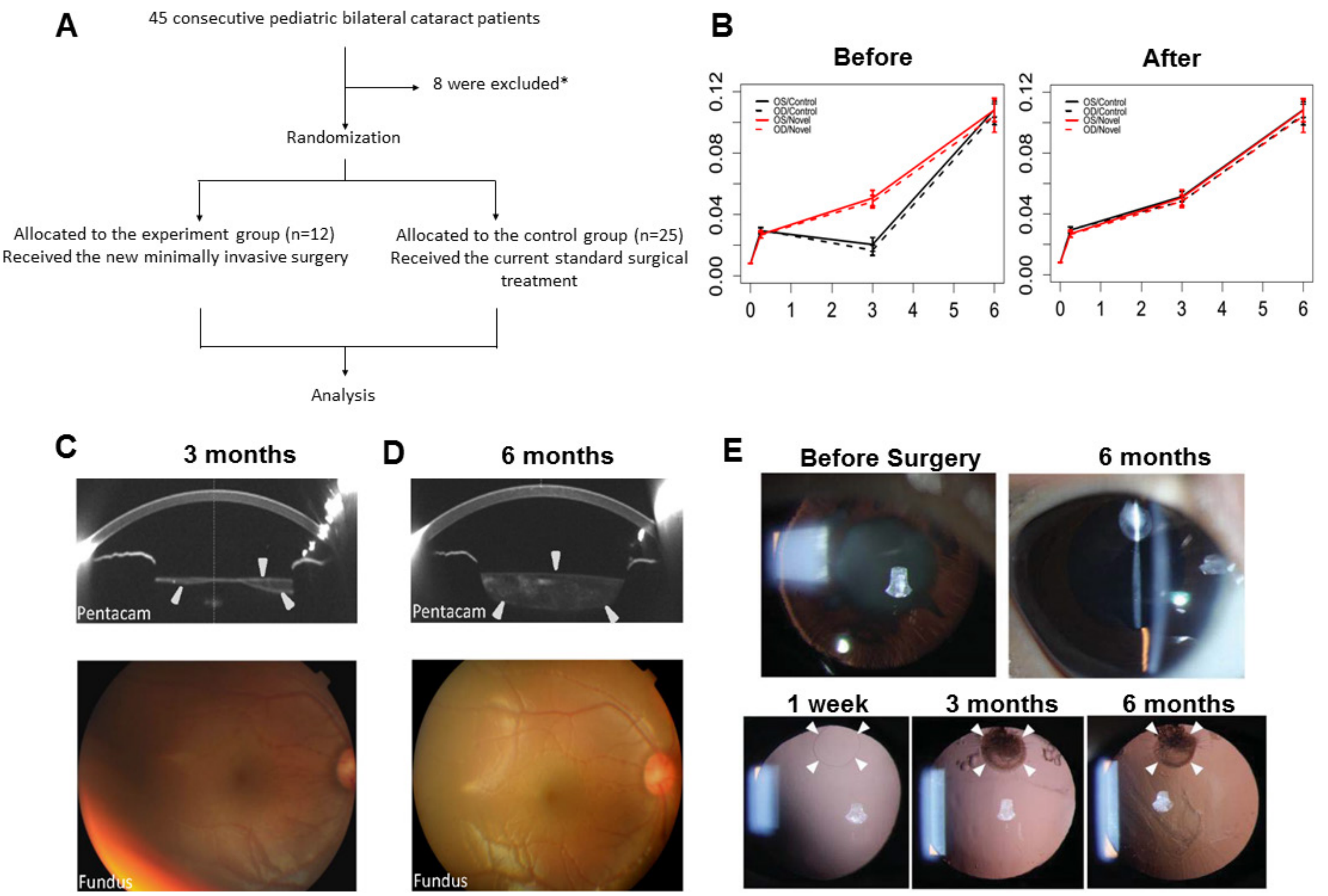
Extended Data Figure 5 | Higher expression levels of *Bmi1*, *Sox2* and *Ki67* in Pax6^{+} LECs. a, Pax6–GFP^{+} LECs were observed at the germinative zone. Left panel, a section of lens of a *Pax6P0-3.9-GFPcre* mouse at P1. Middle and right panels, higher magnification of the framed area in the left panel. Blue indicates DAPI staining. **b**, Upper panel, bright-field photograph showing flat-mount preparation of a lens capsule of a *Pax6P0-3.9-GFPcre* mouse at 6 months; lens capsule materials between two red circles were dissected to enrich Pax6–GFP^{+} LECs. Lower panel, fluorescence image of GFP^{+} LECs from the framed area in the upper panel. AC, anterior capsule; PC, posterior capsule. **c**, Comparison of gene expression levels in Pax6–GFP^{+} LECs versus GFP^{-} LECs in anterior lens capsule in 6-month-old mice, increased expression of the following genes were observed: 10.1-fold in *Pax6* ($P < 0.005$), 8.2-fold in *Ki67* ($P < 0.05$), 4.3-fold in *Bmi1* ($P < 0.05$), and 2.6-fold in *Sox2* ($P < 0.05$), all $n = 5$. Data shown as mean ± s.d.



Extended Data Figure 6 | Lens regeneration surgery in rabbits. **a**, A 3.2-mm keratome was used to make a limbus tunnel incision at the 11–12 o'clock position into the anterior chamber. **b**, The capsular opening was created by a capsulorhexis needle. **c**, A 1–2 mm diameter anterior capsulotomy was performed using the anterior continuous curvilinear capsulorhexis (ACCC) technique near the capsular opening area (yellow arrow). **d**, A blunt needle was used to inject balanced salt solution for hydrodissection of the cortex from the anterior capsule. **e**, The cortex was removed using a phacoemulsification device. **f**, The remaining cortex was removed using irrigation and aspiration. **h**, An elbow I/A handle was used to clear the equatorial cortex. **i**, **j**, The limbus wound was sutured with an interrupted 10-0 nylon suture. The wound was found to be watertight.



Extended Data Figure 7 | Lens regeneration in rabbits. a, Haematoxylin and eosin staining of regenerated lenses at different time points after surgery. At postoperative day 1, a monolayer of LECs between the anterior and posterior capsules was visible (arrowheads). At postoperative day 4, LECs proliferated and covered the posterior capsule. At postoperative day 7, LECs in the posterior capsule began to elongate and differentiate. **b**, At postoperative day 28, LECs in the posterior capsule further elongated, forming primary lens fibres. **c**, Transparency and shape of regenerated lenses in rabbits. Upper panel, slit-lamp photography of a regenerated lens at different time points after surgery. Lower panel, schematic diagram of slit-lamp photographs in the upper panel. At day 1 after surgery, the capsular opening was clearly seen in the peripheral anterior capsule, and the area of LEC loss during surgery is indicated. At 7 weeks after surgery, loss of LECs led to adhesion between the anterior and the posterior capsule and inhibition of lens regeneration in this area.



Extended Data Figure 8 | Human lens regeneration. a, A clinical trial consort flowchart. **b**, Comparison of visual acuity mean response profiles in two groups. A non-parallel pattern of mean responses between two groups was observed largely due to the vision loss at 3 months before laser surgery in the control group (left panel), whereas a parallel pattern of mean responses between two groups was observed using time points including 3 months after laser surgery (right panel); $n = 25$ control, $n = 12$ experimental. Data are shown as mean ± s.d. **c**, Lens thickness increased after surgery. Pentacam showed that 3 months after surgery, the regenerating lens tissue grew from the periphery of the capsular bag to the centre. The sealed capsular bag was only partially filled, appearing spindle-shaped on cross-sectional scan. The fundus was clearly visible on ophthalmoscopy. Arrowheads indicate the regenerated lens structure. **d**, Six months after surgery, the capsular bag was filled with regenerated lens tissue and appeared biconvex on cross-sectional scan by Pentacam. The anterior–posterior capsular adhesion disappeared. The fundus could be seen clearly using an ophthalmoscope with an 18-dioptre lens. **e**, Minimally invasive capsulorhexis preserved LECs for lens regeneration in human infants. Top panel, slit-lamp exam demonstrating human infant's eye visual axis transparency 6 months after minimally invasive surgery compared to baseline (before cataract surgery). Bottom panel, retro-illumination demonstrating the reduced size of the capsulorhexis (white arrowheads).

Extended Data Table 1 | Clinical Outcome Analysis

Extended Data Table 1a. Linear mixed-effect model with decimal acuity as outcome; time, treatment and their interaction as fixed effects; and patient as random effect.

baseline, 1 week, 3 months after surgery and 6 months					baseline, 1 week, 3 months before surgery and 6 months				
	Estimate	Std.Error	Z test	Pr(>|Z|)		Estimate	Std.Error	Z test	Pr(>|Z|)
(Intercept)	0.008	0.003	2.97	0.003**	(Intercept)	0.008	0.003	2.781	0.005**
1 week	0.022	0.003	6.926	<.001***	1 week	0.022	0.003	6.859	<.001***
3 months	0.042	0.003	13.409	<.001***	3 months	0.011	0.003	3.35	<.001***
6 months	0.099	0.003	31.722	<.001***	6 months	0.099	0.003	31.417	<.001***
Trmt (Novel)	0	0.005	-0.024	0.981	Trmt(Novel)	0	0.005	-0.023	0.982
1 week*Trmt	-0.003	0.005	-0.494	0.621	1 week*Trmt	-0.003	0.006	-0.49	0.624
3 months*Trmt	0	0.005	-0.036	0.971	3 months*Trmt	0.031	0.006	5.619	<.001***
6 months*Trmt	-0.001	0.005	-0.093	0.926	6 months*Trmt	-0.001	0.006	-0.092	0.927
Random effect	0.008				Random effect	0.01			
-2logL	1509.948				-2logL	1497.237			

Extended Data Table 1b. Linear mixed-effect model with decimal acuity as outcome; time and treatment as fixed effect; and patient as random effect.

baseline, 1 week, 3 months after surgery and 6 months					baseline, 1 week, 3 months before surgery and 6 months				
	Estimate	Std.Error	Z test	Pr(>|Z|)		Estimate	Std.Error	Z test	Pr(>|Z|)
(Intercept)	0.008	0.003	3.346	<.001***	(Intercept)	0.006	0.003	2.107	0.035*
1 week	0.021	0.003	8.126	<.001***	1 week	0.021	0.003	7.347	<.001***
3 months	0.042	0.003	16.374	<.001***	3 months	0.021	0.003	7.313	<.001***
6 months	0.099	0.003	38.731	<.001***	6 months	0.099	0.003	35.017	<.001***
Trmt(Novel)	-0.001	0.004	-0.276	0.783	Trmt(Novel)	0.007	0.004	1.746	0.081
Random effect	0.008				Random effect	0.009			
-2logL	1536.077				-2logL	1476.647			

Extended Data Table 1c. Likelihood ratio test of fixed effects based on the analysis of response profiles

baseline, 1 week, 3 months after surgery and 6 months				baseline, 1 week, 3 months before surgery and 6 months			
	DF	Chi-Squared	P-value		DF	Chi-Squared	P-value
Time*Treatment	3	0.322	0.956	Time*Treatment		47.529	<.001***
Time	3	532.308	<.001***	Time	3	495.562	<.001***
Treatment	1	0.081	0.776	Treatment	1	3.089	0.079

a, **b**, Linear mixed-effects model (decimal acuity as outcome; time, treatment and their interaction as fixed effects; and patient as random effect). **c**, Likelihood ratio test of fixed effects based on the analysis of response profiles.

ARTICLE

doi:10.1038/nature16966

Observing cellulose biosynthesis and membrane translocation *in crystallo*

Jacob L. W. Morgan[1]*, Joshua T. McNamara[1]*, Michael Fischer[2]†, Jamie Rich[2]†, Hong-Ming Chen[2], Stephen G. Withers[2] & Jochen Zimmer[1]

Many biopolymers, including polysaccharides, must be translocated across at least one membrane to reach their site of biological function. Cellulose is a linear glucose polymer synthesized and secreted by a membrane-integrated cellulose synthase. Here, *in crystallo* enzymology with the catalytically active bacterial cellulose synthase BcsA–BcsB complex reveals structural snapshots of a complete cellulose biosynthesis cycle, from substrate binding to polymer translocation. Substrate- and product-bound structures of BcsA provide the basis for substrate recognition and demonstrate the stepwise elongation of cellulose. Furthermore, the structural snapshots show that BcsA translocates cellulose via a ratcheting mechanism involving a 'finger helix' that contacts the polymer's terminal glucose. Cooperating with BcsA's gating loop, the finger helix moves 'up' and 'down' in response to substrate binding and polymer elongation, respectively, thereby pushing the elongated polymer into BcsA's transmembrane channel. This mechanism is validated experimentally by tethering BcsA's finger helix, which inhibits polymer translocation but not elongation.

Cellulose is an abundant structural cell component produced by many organisms, including bacteria, vascular plants and animals[1–4]. It is a linear polymer of glucose molecules joined between their C1 and C4 carbons[5]. Cellulose is synthesized by membrane-integrated glycosyltransferases (GTs) that contain 6 to 8 transmembrane helices (TMHs) as well as an intracellular catalytic GT domain[6]. These enzymes polymerize UDP-activated glucose (UDP-Glc)[7,8] into chains thousands of glucose units long[9] and translocate the polymer across the plasma membrane, through a pore formed by their own transmembrane region[10].

Cellulose is also a common biofilm component[2,11] where it is synthesized and secreted via an inner and, in Gram-negative bacteria, outer membrane-spanning cellulose synthase complex[3]. At the inner membrane, the catalytic BcsA and membrane-anchored, periplasmic BcsB subunits form a complex sufficient to synthesize and translocate cellulose[7], while transport across the outer membrane probably occurs through the BcsC subunit[12,13].

Processive glycosyltransferases, including chitin, alginate and cellulose synthases, transfer the glycosyl moiety from a nucleotide-activated sugar (donor) to a specific hydroxyl group of the growing polysaccharide chain (acceptor) by a nucleophilic S_N2-like substitution reaction[14], thereby forming an elongated polymer and nucleoside diphosphate as reaction products. A processive mechanism requires that the elongated polymer is translocated after each glycosyl transfer, such that the polymer's newly added sugar unit becomes the acceptor in a subsequent reaction. Because all known processive glycosyltransferases are transmembrane channel-forming enzymes[15–17], the translocation of the polymer into the transmembrane channel between catalytic steps also gives rise to secretion.

Previous structural and functional analyses of the *Rhodobacter sphaeroides* BcsA–BcsB complex containing a nascent cellulose polymer revealed the architecture of the active site, its close association with the transmembrane channel, as well as the coordination of cellulose within the channel[10]. In bacteria, cellulose biosynthesis is activated by the signalling molecule cyclic-di-GMP (c-di-GMP)[18], a potent biofilm inducer and allosteric activator of BcsA[19]. Binding of the activator to BcsA's carboxy-terminal PilZ domain allows a "gating loop" to either insert into the catalytic pocket during substrate binding or to retract from it to release the UDP product[20].

Cellulose synthases contain a short helix within the GT domain, termed "finger helix"[20]. The amino terminus of the finger helix contacts the polymer's acceptor glucose via an invariant "TED" motif, of which the aspartic acid (D) probably facilitates the deprotonation of the acceptor C4 hydroxyl during catalysis[10,20].

Crystal structures of the catalytically inactive "resting" state of BcsA–BcsB (in the absence of c-di-GMP)[10] and a c-di-GMP-activated structure[20] provided important insights into the architecture and function of processive glycosyltransferases. Here we used *in crystallo* enzymology to obtain structural snapshots of a complete cellulose biosynthesis reaction cycle, providing structures of substrate- and product-bound states, and delineating the mechanism by which the elongated glucan is translocated into BcsA's transmembrane channel.

BcsA elongates the chain one glucose unit at a time

The previously determined c-di-GMP-activated BcsA–BcsB structure[20] contains a nascent cellulose polymer 18 glucose molecules long whose non-reducing terminal glucose unit rests at the entrance to BcsA's transmembrane channel, which is marked by the invariant Trp383 of the "QxxRW" motif[20,21]. In this state, BcsA's finger helix is in an 'up' position where it points towards the entrance of the transmembrane channel, thereby positioning Asp343, the putative catalytic base, near the C4 hydroxyl of the polymer's terminal sugar (Fig. 1a). Because BcsA's active site is empty and the gating loop is retracted from it, this structure represents a state in which the enzyme is poised to initiate a new cycle of chain elongation, hereafter referred to as the 'post-translocation state'.

Notably, BcsA–BcsB is catalytically active *in crystallo*. Incubating BcsA–BcsB crystals with UDP-Glc in the absence of Mg^{2+} (to slow

[1]University of Virginia School of Medicine, Center for Membrane Biology, Molecular Physiology and Biological Physics, 480 Ray C. Hunt Drive, Charlottesville, Virginia 22908, USA. [2]Department of Chemistry, University of British Columbia, 2036 Main Mall, Vancouver, British Columbia V6T 1Z1, Canada. †Present Addresses: Sandoz GmbH, Biochemiestrasse 10, A-6250 Kundl, Austria (M.F.); Kairos Therapeutics, 2405 Wesbrook Mall, Fourth Floor, Vancouver, British Columbia V6T 1Z3, Canada (J.R.).
*These authors contributed equally to this work.

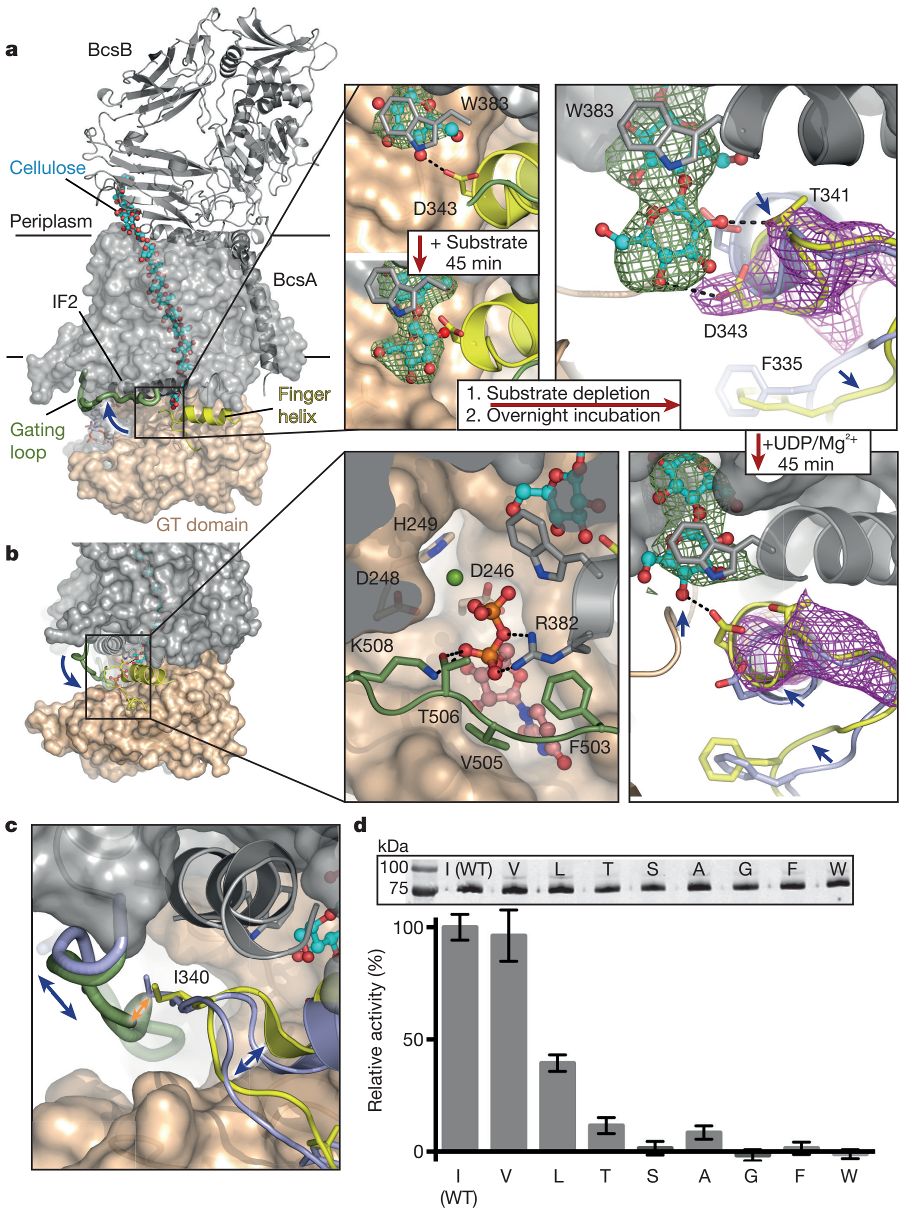


Figure 1 | *In crystallo* cellulose biosynthesis. BcsA is shown as a surface with the transmembrane and GT regions coloured grey and beige, respectively. BcsB is shown as a grey cartoon. The cellulose polymer is shown as cyan sticks, BcsA's finger helix and gating loop are shown as a cartoon coloured yellow and green, respectively. Trp383 is represented in grey sticks as a marker for the transmembrane channel entrance. **a**, Left, organization of the BcsA–BcsB complex and formation of a channel for the translocating polymer. Right, the pre-translocation state of BcsA. Unbiased Sigma-A weighted $F_o - F_c$ difference electron densities contoured at 4 and 3σ are shown as green and magenta meshes for the nascent cellulose chain and finger helix, respectively. **b**, Translocation of cellulose. Crystals described in **a** were subsequently soaked with UDP/Mg^{2+}. UDP is shown in sticks and coloured violet for the carbon atoms and Mg^{2+} is shown as a green sphere. **c**, Insertion of the gating loop into the active site is incompatible with the 'down' position of the finger helix, probably owing to a clash between Ile340 and the gating loop's backbone. The retracted gating loop and finger helix in the 'down' position are both shown as blue cartoons. **d**, Cellulose biosynthesis by BcsA I340 mutants. Ile340 was replaced with the indicated residues and *in vitro* cellulose biosynthesis was performed in IMVs as described[7]. All activities are represented relative to the wild-type (WT) activity, and error bars represent standard deviations from 3 replicates. Inset, western blot analysis of the IMVs used, showing equal expression levels of all BcsA mutants.

down the reaction) results in the extension of the polymer's electron density by one glucose unit (Fig. 1a). This demonstrates that cellulose elongation occurs via a stepwise addition of glucose units and that Trp383 of the QxxRW motif indeed forms the acceptor-binding site. The elongated polymer points straight into the catalytic pocket, similar to its position in the recently determined resting state of BcsA[10].

BcsA's finger helix resets upon polymer extension

Processive cellulose biosynthesis requires that the elongated polymer is translocated after each elongation cycle and the above described *in crystallo* cellulose extension demonstrates that glycosyl transfer and polymer translocation are separate steps.

To identify whether the extended cellulose translocates spontaneously over time, we extended the polymer *in crystallo* as described above, then diluted the substrate 65-fold, and incubated the crystals overnight before harvesting. Under these conditions, the density for the extended polymer continues to protrude into the catalytic pocket, suggesting that this state is stable in the absence of substrate (Fig. 1a). Notably, after extending the cellulose polymer, BcsA's finger helix shifts to a 'down' position, such that Thr341 and Asp343 of its TED motif again form hydrogen bonds with the polymer's terminal glucose unit (Fig. 1a).

The finger helix movement is accompanied by retraction of a small loop, which contains Phe335, from a hydrophobic pocket beneath the finger helix (Fig. 1a). The resulting 'pre-translocation' state of BcsA–BcsB contains an empty catalytic pocket, a gating loop retracted from the active site (Extended Data Fig. 1), an extended polymer, and a downward pointing finger helix (Fig. 1a).

Coupled movement of gating loop and finger helix

The observed downward movement of BcsA's finger helix in response to cellulose elongation suggests its role in translocation. If crystals containing the above described pre-translocation step are soaked with UDP/Mg^{2+}, UDP binds and the gating loop inserts into the catalytic pocket (Fig. 1b). Additionally, the finger helix returns to the 'up' position and the density for the polymer's newly-added glucose unit disappears (Fig. 1b), suggesting its translocation into the channel.

To confirm that the elongated polymer is indeed translocated *in crystallo*, we extended the polymer by one unit with the chain-terminating analogue 6-thio-galactose, whose location can be unambiguously determined based on anomalous X-ray scattering of its sulfur atom. After polymer extension, the 6-thio-galactosyl moiety sits inside BcsA's catalytic pocket; then upon UDP/Mg^{2+} binding and gating loop insertion into the active site, the density of the newly-added sugar

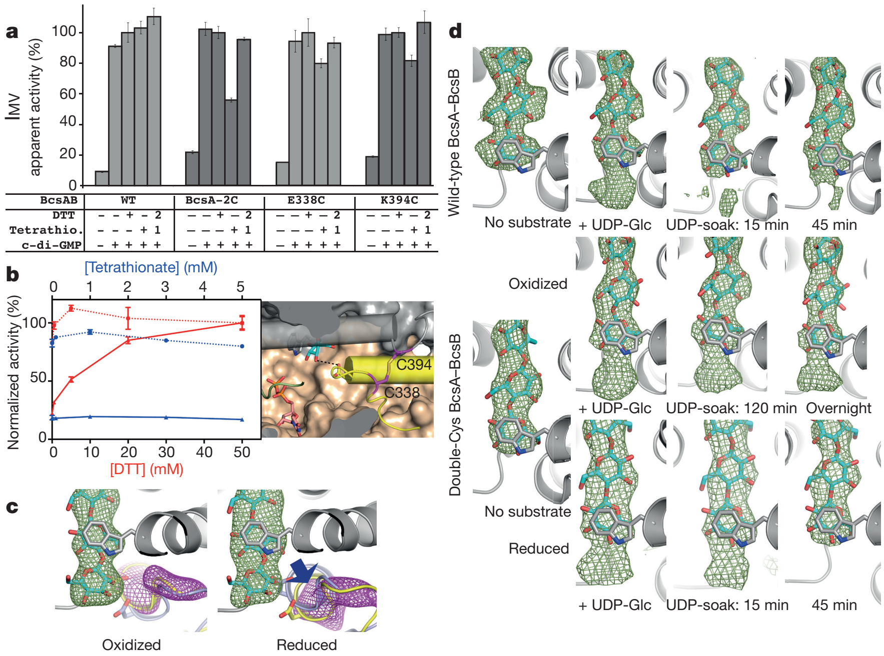


Figure 2 | Movement of BcsA's finger helix is essential for cellulose translocation. *In vitro* cellulose formation by wild-type and mutant BcsA–BcsB complexes. **a**, Activity in IMVs under reducing (DTT) and oxidizing (sodium tetrathionate) conditions. For each mutant tested, the apparent activity was normalized to its activity in the presence of 50 mM DTT and (1) and (2) indicate the order of DTT and tetrathionate addition. **b**, DTT and tetrathionate titrations are shown in red and blue, respectively, for wild-type BcsA (dashed lines) and BcsA-2C (solid lines). All experiments were performed in triplicate and error bars represent the deviations from the means. Right, location of the engineered disulfide bond in BcsA-2C, coloured as in Fig. 1. **c**, Comparison of BcsA-2C finger helix positions following polymer extension under oxidizing and reducing conditions. Upon cellulose elongation, crystals were oxidized or reduced and incubated without substrate for 16 h or 6 h, respectively. **d**, Comparison of *in crystallo* cellulose translocation in wild type and BcsA-2C. The cellulose polymer was extended, then translocation was initiated as described in Fig. 1, and crystals were harvested after the indicated incubation periods. In all panels, the unbiased electron densities for the glucan and finger helix are contoured and coloured as described in Fig. 1. Trp383 of the acceptor-binding site is shown in grey.

disappears, and the thio-galactosyl unit moves into the transmembrane channel next to Trp383 (Extended Data Fig. 2), thereby confirming the genuine translocation of cellulose *in crystallo*.

The ability of UDP/Mg^{2+} to induce translocation brings about the question of why translocation doesn't occur immediately after glycosyl transfer when the UDP/Mg^{2+} product is bound at the active site and the gating loop is inserted. All states of BcsA observed thus far either show its gating loop inserted into the active site and the finger helix in the 'up' position (Fig. 1b), or the gating loop retracted from the active site and the finger helix in the 'down' position, if the cellulose polymer is extended[10,20] (Fig. 1a). This suggests that the finger helix cannot move to the 'down' position unless the gating loop is retracted from the active site and that in turn gating loop insertion could induce the upward movement of the finger helix. This coupled movement is probably due to steric clashes between the side chain of Ile340, preceding the TED motif of the finger helix, and the gating loop's backbone (Fig. 1c). Ile340 is primarily conserved among bacterial cellulose synthases, eukaryotic enzymes usually contain a valine or occasionally a leucine residue at this position, which could perform a similar function. Indeed, BcsA carrying a Val at position 340 shows indistinguishable catalytic activity compared to the wild-type enzyme, while an Ile to Leu substitution reduces the apparent activity by about 50% (Fig. 1d). Thr, Ala, Ser, Phe or Trp, however, support only low or background activities, suggesting that gating loop to finger helix coupling requires a fairly rigid, hydrophobic residue at position 340 (Fig. 1d).

Finger helix must move for cellulose translocation

The conformational changes of BcsA described above suggest that its finger helix moves up and down during cellulose translocation. To test this hypothesis, we engineered double-Cys BcsA mutants expected to crosslink the finger helix to an amphipathic helix (IF2) above the GT domain[10] (Fig. 1a and Extended Data Fig. 3), and screened those mutants for changes in catalytic activity upon oxidation. Introducing Cys residues at positions 338 near the N terminus of the finger helix and 394 at the C-terminal end of IF2 (hereafter referred to as BcsA-2C) results in attenuated catalytic activity compared to the wild-type enzyme or single-Cys mutants in inverted membrane vesicles (IMVs). Full activity, however, can be restored upon addition of excess DTT (Fig. 2a). Upon purification of BcsA-2C and reconstitution into proteoliposomes, its catalytic activity further decreases to ~20% under non-reducing conditions (Fig. 2b), probably owing to complete disulfide bond formation during purification. Addition of the oxidizing reagent tetrathionate does not further decrease the enzyme's apparent activity, yet the catalytic activity robustly recovers with increasing DTT concentrations. The crystal structure of this BcsA-2C mutant reveals that the disulfide bond forms when the finger helix is in the 'up' position (Fig. 2b and Extended Data Fig. 3), similar to its position in the post-translocation state[20].

Additionally, the ability to observe cellulose elongation and translocation *in crystallo* allows us to further delineate how the engineered disulfide bond affects BcsA's activity. Soaking BcsA-2C crystals with substrate leads to polymer extension as observed for wild-type BcsA, demonstrating that glycosyl transfer is not abolished by the mutations (Fig. 2c, d). Subsequently, if those crystals are then incubated overnight under oxidizing conditions, the finger helix remains in the 'up' position, revealing that the cross-link prevents the finger helix from resetting to the 'down' position (Fig. 2c). Binding of UDP/Mg^{2+} in these crystals, which induces polymer translocation in wild-type BcsA within 15 to 45 min, fails to initiate translocation, even after an overnight incubation (Fig. 2d). However, reducing the engineered disulfide bond in the BcsA-2C complex restores the capability of the finger helix to move downwards following polymer extension and, most notably, restores polymer translocation (Fig. 2c, d), thereby directly correlating the movement of the finger helix with BcsA's ability to translocate the polymer.

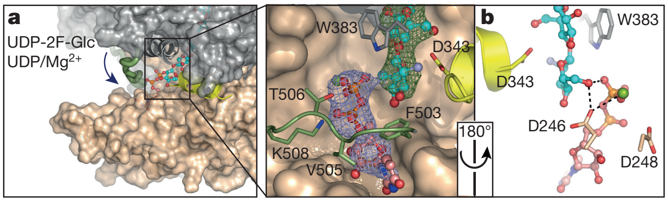


Figure 3 | The product-bound state. The product-bound state of BcsA contains an elongated cellulose polymer and an inserted gating loop coordinating UDP/Mg^{2+} at the active site. **a**, Cellulose was elongated *in crystallo* with a 2-deoxy-2-fluoro-glucose moiety and UDP/Mg^{2+} was rebound to the active site as described in Fig. 1. Unbiased Sigma-A weighted $F_o - F_c$ difference electron densities contoured at 4 and 4.5σ are shown for the nascent cellulose polymer and UDP/Mg^{2+} in green and blue, respectively. **b**, The terminal glucose unit of the extended cellulose polymer forms interactions with the β-phosphate of UDP and Asp246 of the DxD motif. Colours are as in Fig. 1 and fluorine is shown in light blue.

The product-bound state

Directly after glycosyl transfer, BcsA contains an elongated glucan plus UDP/Mg^{2+} and an inserted gating loop at the active site, as well as the finger helix in the 'up' position. This 'product-bound' state is accessible through the BcsA-2C mutant described above. Because the engineered disulfide bond tethers the finger helix, UDP/Mg^{2+} can be bound to the active site after polymer extension without inducing translocation.

Alternatively, we observed that incorporating 2-fluoro-substituted glucose into the polymer stabilizes a similar product-bound state. Attempting to trap a 'donor-bound' state by using the usually (but not always) unreactive UDP-2-fluoro-glucose as substrate[22,23], we observed that the nascent glucan is elongated. However, the subsequent translocation of this polymer is significantly impeded (but not abolished), which may be due to the loss of a hydrogen bond between Thr341 of the finger helix and the polymer's terminal C2 substituent, Fig. 1a. Thus, elongating the cellulose polymer with 2-fluoro-glucose in wild-type BcsA, then diluting the substrate and soaking in UDP/Mg^{2+} reproduces a similar product-bound state to that obtained for the BcsA-2C mutant. Due to higher quality diffraction data (Extended Data Table 1), we discuss the structure obtained with 2-fluoro-glucose.

The product-bound BcsA structure shows that the catalytic pocket can accommodate both an extended cellulose polymer and UDP/Mg^{2+}, suggesting that the newly added glucose unit can align with the polymer before the gating loop retracts from the active site and UDP is released (Fig. 3). In this position, the only major interactions of the terminal glucose unit are with UDP's β-phosphate as well as Asp246 of the 'DxD' motif[6,14] via its C2 or C6 hydroxyl group (depending on its orientation) (Fig. 3b).

Of note, the individual glucose units in cellulose are rotated by approximately 180° relative to their neighbours[5,24]. Therefore, during relaxation into the polymer's plane, the newly added glucose moiety must rotate either clockwise or counter clockwise to be in register with the preceding glucose units[25,26]. This alternating rotation is most likely to be driven by the formation of intramolecular hydrogen bonds[26] and could be facilitated by UDP release after glycosyl transfer, which minimizes steric restrictions at the active site (Fig. 3).

The donor glucose binds to a pocket under the acceptor

To stabilize the substrate-bound state of BcsA, we synthesized and employed a non-hydrolysable phosphonate substrate analogue[27], (UDP-CH_2-Glc), in which a methylene bridge connects the donor glucose with UDP's β-phosphate. Additionally, we also capped the cellulose polymer with galactose (see Methods), which cannot be extended owing to an axial instead of an equatorial hydroxyl group at its C4 position.

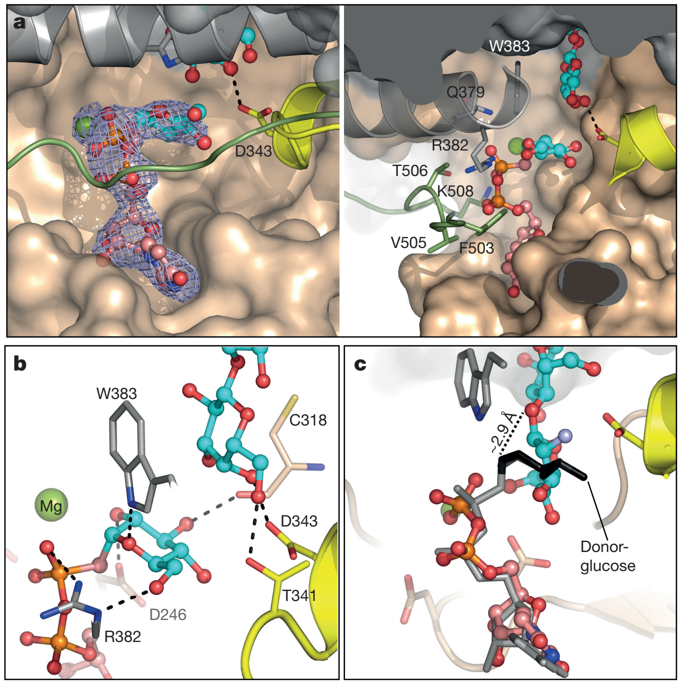


Figure 4 | The substrate-bound state. The donor glucose binds in a conserved pocket beneath the acceptor. **a**, Unbiased Sigma-A weighted $F_o - F_c$ difference electron density of UDP-CH_2-glucose at BcsA's active site, contoured at 4σ. Right, conserved residues of BcsA involved in coordinating the donor glucose are shown as sticks. **b**, Interactions between the donor sugar moiety and BcsA's D_{246}xD, FFC_{318}GS, $T_{341}ED_{343}$ and $QxxR_{382}W_{383}$ motifs. **c**, Comparison of substrate- and product-bound states. Aligning the substrate's pyrophosphate group with the position of UDP in the product-bound state (Fig. 3) positions the donor's C1 carbon within approximately 2.9 Å of the acceptor's hydroxyl group. The substrate is shown as grey and black sticks for its UDP and glucose moieties, respectively.

The donor glucose inserts underneath the acceptor into a conserved hydrophilic pocket formed by BcsA's TED, HAKAG and FFCGS motifs (Fig. 4a), and the gating loop cooperates with the QxxRW motif to stabilize the substrate. In particular, the gating loop's Phe503 forms cation-π interactions with Arg382 of the QxxRW motif, which in turn forms a salt bridge with the nucleotide's β-phosphate as well as a hydrogen bond with the donor's C6 hydroxyl (Fig. 4a, b). Further, the donor's C3 hydroxyl interacts with the backbone carbonyl of Cys318 of the FFCGS motif and its ring oxygen is in hydrogen bond distance to the Nε of Trp383 of the QxxRW motif (Fig. 4b). Thus, the recognition of the donor's C3 and C6 hydroxyls and its ring oxygen seems to be particularly important for substrate selectivity.

In the substrate-bound state, BcsA's finger helix is in the 'up' conformation and positions Asp343 of the TED motif within 2.5 Å of the acceptor's C4 hydroxyl group, consistent with its likely function as general base during catalysis (Fig. 4b). However, the distance between the acceptor and the donor's C1 carbon is about 4.2 Å (assuming glucose instead of galactose as the polymer's terminal sugar), which is probably too far for a direct transfer. We note that the pyrophosphate group of UDP-CH_2-Glc is less deeply inserted into the active site compared to UDP in the product-bound state (Extended Data Fig. 4), perhaps owing to the substrate's methylene-bridge and/or the capping of the glucan with galactose. Repositioning UDP-CH_2-Glc according to the UDP conformation in the product-bound state places the donor's C1 carbon within approximately 2.9 Å of the acceptor's C4 hydroxyl, a suitable distance for glycosyl transfer[28] (Fig. 4c). This distance is likely to be also maintained when the acceptor is in the opposite orientation (as is the case for every other glucose unit), owing to repositioning of the terminal glucose unit at the active site.

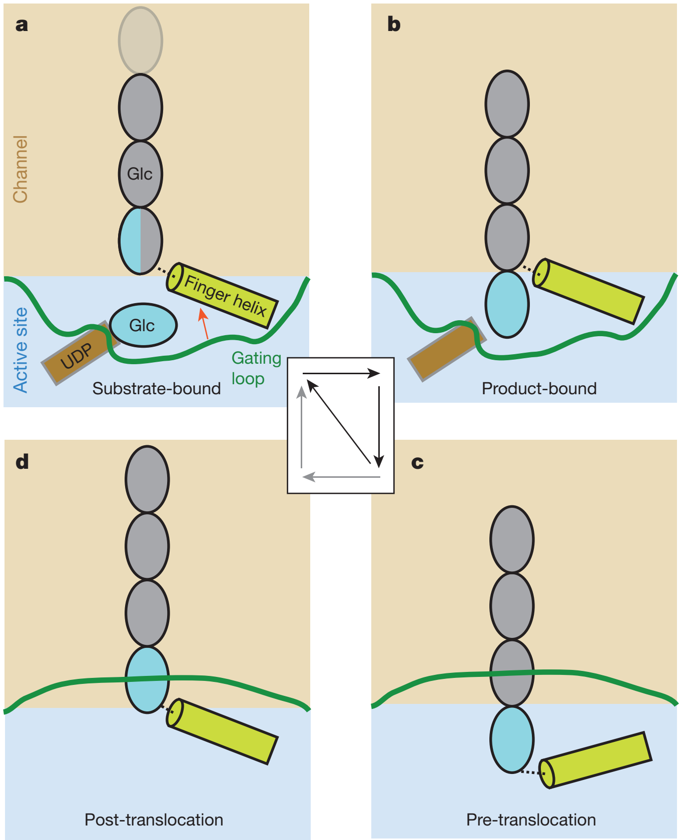


Figure 5 | Model of cellulose biosynthesis. a, Cellulose biosynthesis might start with substrate binding to BcsA when the polymer's terminal glucose unit sits at the acceptor site at the entrance to BcsA's transmembrane channel. At this time, the gating loop stabilizes the substrate and the finger helix in the 'up' position (red arrow). **b**, **c**, Glycosyl transfer generates the product-bound state (**b**) and retraction of the gating loop and UDP release allows the finger helix to reset to the 'down' position to contact again the polymer's terminal glucose unit (**c**). Substrate binding to this pre-translocation state and insertion of the gating loop could induce the upward movement of the finger helix and polymer translocation (**a**) or spontaneous translocation might precede substrate binding via a post-translocation state (**d**).

Implications for cellulose biosynthesis

Cellulose biosynthesis requires that BcsA binds the substrate, positions it for and catalyses glycosyl transfer, translocates the extended polymer, and exchanges UDP with UDP-Glc for a subsequent elongation cycle. Our structural snapshots of a complete cellulose biosynthesis cycle suggest that BcsA accomplishes this in three steps (Fig. 5).

Upon substrate binding, BcsA's gating loop inserts into the catalytic pocket, thereby positioning the donor glucose for transfer and perhaps also stabilizing the UDP leaving group (Fig. 5). In this state, the acceptor glucose rests next to Trp383 at the entrance to the transmembrane channel and interacts with the TED motif at the N terminus of the finger helix. After glycosyl transfer, the newly added glucose unit aligns with the polymer and extends into the catalytic pocket next to UDP's pyrophosphate group. In this product-bound state, the gating loop remains inserted into the active site and the finger helix continues to point 'up' as observed in the substrate-bound state. Next, BcsA's gating loop retracts to release UDP and the finger helix resets to the 'down' position to interact again with the polymer's terminal glucose unit. Binding of a new substrate molecule to this pre-translocation state could elicit the translocation of the extended polymer (by an upward movement of the finger helix) through re-insertion of the gating loop into the active site. *In crystallo* translocation experiments with a galactose-capped polymer and UDP-Glc as substrate confirmed that the polymer can indeed be translocated when UDP-Glc/Mg^{2+} binds to the active site (Extended Data Fig. 5). An alternative, perhaps slower, pathway could be the translocation of the polymer before substrate binding, facilitated either by random cycles of gating loop insertion and retraction or favourable interactions of the polymer in the extracellular milieu.

How could the opening of the gating loop and resetting of the finger helix be coordinated? The retraction of the gating loop after glycosyl transfer could be facilitated by the rotation of the newly added sugar into the plane of the polymer. Alternatively, it is possible that the Mg^{2+} coordination changes after glycosyl transfer, which in turn could affect the stability of UDP and the gating loop at the active site, as proposed for non-processive galactosyl transferases[29].

BcsA's finger helix is capped at its N terminus by the TED motif, which is invariant among cellulose synthases. Within the motif, Thr341 and Asp343 form hydrogen bonds with the polymer's terminal sugar unit, which may enable the finger helix to exert force on the polymer during the upward movement. Both residues are well-suited for this task: as a β-branched amino acid, the side chain hydroxyl of Thr is sterically restricted, and Asp343 is further rigidified by interactions with its backbone amide proton and that of the following residue (Extended Data Fig. 6).

How might the finger helix move downward without retro-translocating the glucan? N-terminal capping of α-helices with Asp residues has been shown to significantly stabilize the helical conformation in a pH dependent manner[30,31]. During catalysis, Asp343 abstracts a proton from the acceptor's C4 hydroxyl group[14], thereby probably altering its interaction with the amide protons at the N terminus of the finger helix. Thus, we speculate that the destabilization of the finger helix during glycosyl transfer enables it to refold in the 'down' position, after UDP release and subsequent deprotonation of Asp343. This notion is supported by the position of another conserved residue within the finger helix. Thr346 sits at the membrane distal side of the finger helix just three residues C-terminal of Asp343, and its side chain hydroxyl provides an alternative hydrogen bond partner for the backbone carbonyl of the preceding Glu342 (Extended Data Fig. 6). Threonine residues in α-helices, in particular in hydrophobic environments, often induce helical kinks, which could facilitate the repositioning of the finger helix[32,33].

On its own, the finger helix is insufficient for cellulose translocation, which requires substrate or UDP binding and gating loop insertion into the active site. Thus, we conclude that the free energy of substrate binding energizes cellulose translocation. Additional thermodynamic driving force for translocation may be generated by the base catalyst itself. The post-translocation state is likely to be energetically favourable, owing to a strong interaction between Asp343 and the acceptor's C4 hydroxyl. This interaction is broken upon protonation of its side chain during catalysis but re-established after polymer translocation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 June 2015; accepted 5 January 2016.
Published online 9 March 2016.

1. Keegstra, K. Plant cell walls. *Plant Physiol.* **154,** 483–486 (2010).
2. Serra, D. O., Richter, A. M. & Hengge, R. Cellulose as an architectural element in spatially structured *Escherichia coli* biofilms. *J. Bacteriol.* **195,** 5540–5554 (2013).
3. Römling, U. Molecular biology of cellulose production in bacteria. *Res. Microbiol.* **153,** 205–212 (2002).
4. Kimura, S., Ohshima, C., Hirose, E., Nishikawa, J. & Itoh, T. Cellulose in the house of the appendicularian *Oikopleura rufescens*. *Protoplasma* **216,** 71–74 (2001).
5. Nishiyama, Y., Sugiyama, J., Chanzy, H. & Langan, P. Crystal structure and hydrogen bonding system in cellulose Iα from synchrotron X-ray and neutron fiber diffraction. *J. Am. Chem. Soc.* **125,** 14300–14306 (2003).

6. McNamara, J., Morgan, J. L. W. & Zimmer, J. A molecular description of cellulose biosynthesis. *Annu. Rev. Biochem.* **84,** 895–921 (2015).
7. Omadjela, O. *et al.* BcsA and BcsB form the catalytically active core of bacterial cellulose synthase sufficient for *in vitro* cellulose synthesis. *Proc. Natl Acad. Sci. USA* **110,** 17856–17861 (2013).
8. Brown, C., Leijon, F. & Bulone, V. Radiometric and spectrophotometric *in vitro* assays of glycosyltransferases involved in plant cell wall carbohydrate biosynthesis. *Nature Protocols* **7,** 1634–1650 (2012).
9. Somerville, C. Cellulose synthesis in higher plants. *Annu. Rev. Cell Dev. Biol.* **22,** 53–78 (2006).
10. Morgan, J. L., Strumillo, J. & Zimmer, J. Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature* **493,** 181–186 (2013).
11. Bokranz, W., Wang, X., Tschäpe, H. & Römling, U. Expression of cellulose and curli fimbriae by *Escherichia coli* isolated from the gastrointestinal tract. *J. Med. Microbiol.* **54,** 1171–1182 (2005).
12. Whitney, J. C. *et al.* Structural basis for alginate secretion across the bacterial outer membrane. *Proc. Natl Acad. Sci. USA* **108,** 13083–13088 (2011).
13. Keiski, C.-L. *et al.* AlgK is a TPR-containing protein and the periplasmic component of a novel exopolysaccharide secretin. *Structure* **18,** 265–273 (2010).
14. Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.* **77,** 521–555 (2008).
15. Hubbard, C., McNamara, J. T., Azumaya, C., Patel, M. S. & Zimmer, J. The hyaluronan synthase catalyzes the synthesis and membrane translocation of hyaluronan. *J. Mol. Biol.* **418,** 21–31 (2012).
16. Merzendorfer, H. Insect chitin synthases: a review. *J. Comp. Physiol. B* **176,** 1–15 (2006).
17. Rehm, B. H. Alginate production: precursor biosynthesis, polymerization and secretion. *Microbiology Monographs* **13,** 55–71 (2009).
18. Ross, P. *et al.* Regulation of cellulose synthesis in *Acetobacter xylinum* by cyclic diguanylic acid. *Nature* **325,** 279–281 (1987).
19. Römling, U., Galperin, M. Y. & Gomelsky, M. Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. *Microbiol. Mol. Biol. Rev.* **77,** 1–52 (2013).
20. Morgan, J. L. W., McNamara, J. T. & Zimmer, J. Mechanism of activation of bacterial cellulose synthase by cyclic di-GMP. *Nature Struct. Mol. Biol*. **21,** 489–496 (2014).
21. Saxena, I. M., Brown, R. M., Jr & Dandekar, T. Structure–function characterization of cellulose synthase: relationship to other glycosyltransferases. *Phytochemistry* **57,** 1135–1148 (2001).
22. Persson, K. *et al.* Crystal structure of the retaining galactosyltransferase LgtC from *Neisseria meningitidis* in complex with donor and acceptor sugar analogs. *Nature Struct. Mol. Biol*. **8,** 166–175 (2001).
23. Chan, P. H. *et al.* Investigating the structural dynamics of α-1,4-galactosyltransferase C from *Neisseria meningitidis* by nuclear magnetic resonance spectroscopy. *Biochemistry* **52,** 320–332 (2013).
24. Gardner, K. H. & Blackwell, J. The hydrogen bonding in native cellulose. *Biochim. Biophys. Acta* **343,** 232–237 (1974).
25. Yang, H., Zimmer, J., Yingling, Y. G. & Kubicki, J. D. How cellulose elongates–A QM/MM study of the molecular mechanism of cellulose polymerization in bacterial CESA. *J. Phys. Chem. B* **119,** 6525–6535 (2015).
26. Carpita, N. C. Update on mechanisms of plant cell wall biosynthesis: how plants make cellulose and other (1→4)-β-D-glycans. *Plant Physiol.* **155,** 171–184 (2011).
27. Martin, J. L., Johnson, L. N. & Withers, S. G. Comparison of the binding of glucose and glucose 1-phosphate derivatives to T-state glycogen phosphorylase b. *Biochemistry* **29,** 10745–10757 (1990).
28. Vocadlo, D. J., Davies, G. J., Laine, R. & Withers, S. G. Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate. *Nature* **412,** 835–838 (2001).
29. Qasba, P. K., Ramakrishnan, B. & Boeggeman, E. Structure and function of β-1,4-galactosyltransferase. *Curr. Drug Targets* **9,** 292–309 (2008).
30. Forood, B., Feliciano, E. J. & Nambiar, K. P. Stabilization of alpha-helical structures in short peptides via end capping. *Proc. Natl Acad. Sci. USA* **90,** 838–842 (1993).
31. Dirr, H. W., Little, T., Kuhnert, D. C. & Sayed, Y. A conserved N-capping motif contributes significantly to the stabilization and dynamics of the C-terminal region of class alpha glutathione S-transferases. *J. Biol. Chem.* **280,** 19480–19487 (2005).
32. Scharnagl, C. *et al.* Side-chain to main-chain hydrogen bonding controls the intrinsic backbone dynamics of the amyloid precursor protein transmembrane helix. *Biophys. J.* **106,** 1318–1326 (2014).
33. Cao, Z. & Bowie, J. U. Shifting hydrogen bonds may produce flexible transmembrane helices. *Proc. Natl Acad. Sci. USA* **109,** 8121–8126 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Rapoport for critical comments on the manuscript and J. Acheson for advice on reducing BcsA–BcsB complexes *in crystallo*. Diffraction data were collected at the Argonne National Laboratory's Advanced Photon Source (APS) beam lines 23-ID-D (GM/CA-), 22-ID (SER-) and 24-ID-C (NE-CAT). GM/CA@APS has been funded in whole or in part with Federal funds from the National Cancer Institute (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006). The NE-CAT beam lines are funded by the National Institute of General Medical Sciences from the National Institutes of Health (P41 GM103403). The Pilatus 6M detector on 24-ID-C beam line is funded by a NIH-ORIP HEI grant (S10 RR029205). Data for this research was also in part collected at the APS SER-CAT beam line, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by ANL under Contract No. DE-AC02-06CH11357. J.L.W.M. is supported by a National Science Foundation Graduate Research Fellowship, Grant No. DGE-1315231. M.F. thanks the Austrian Science Fund (FWF) (J3293-B21) for an Erwin Schrödinger postdoctoral fellowship. This research was primarily supported by the National Institutes of Health, Grant 1R01GM101001, awarded to J.Z.; S.G.W. thanks the Natural Sciences and Engineering Research Council of Canada for financial support.

Author Contributions J.T.M. and J.L.W.M. purified and crystallized BcsA–BcsB and performed all crystal soaking experiments. J.T.M. cloned and analysed all BcsA cysteine mutants. J.T.M. and J.L.W.M. collected and processed diffraction data and built and refined the BcsA–BcsB models. M.F. synthesized the fluorinated and phosphonate UDP-Glc analogues and J.R. and H.-M.C. synthesized the UDP-thio-galactose analogues. J.T.M., J.L.W.M. and J.Z. analysed the data. J.Z. and J.L.W.M. wrote the paper and all authors edited the text.

Author Information Structure factors and coordinates have been deposited at the Protein Data Bank under entry codes 5EJ1, 5EIY and 5EJZ. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.Z. (jochen_zimmer@virginia.edu).

METHODS

***In crystallo* cellulose synthesis.** *Rhodobacter sphaeroides* BcsA–BcsB was purified and crystallized as previously described[20] by the bicelle crystallization method with the exception that gel filtration and crystallization were carried out in buffers lacking $MgCl_2$. After the crystals grew to full size (about 2 weeks), cryo-protection was initiated by three successive 2 μl additions of cryo solution (well solution containing 20% glycerol) to the crystal mother liquor without added $MgCl_2$, each addition separated by 10 min.

After the third addition of cryo solution, the polymer was elongated by adding 0.8 μl of 100 mM UDP-activated sugar (glucose or 2-fluoro glucose) in the absence of $MgCl_2$ to the ~8 μl crystallization drop for a final UDP-sugar concentration near 10 mM. The crystals were incubated with UDP-sugar for 2–3 h at 30 °C. After incubation, 6 μl of the crystallization solution was replaced with an equal volume of fresh cryo solution and this process was repeated twice to dilute the UDP-sugar concentration approximately 65-fold. Subsequently, crystals were then looped and flash-cooled in liquid N_2 at various time points.

***In crystallo* cellulose translocation.** For *in crystallo* translocation experiments using wild-type BcsA– BcsB, 0.8 μl of a solution containing 100 mM UDP and 250 mM $MgCl_2$ was added to the ~8 μl crystallization drop (see above) for a final concentration of ~10 mM UDP and 25 mM $MgCl_2$. Crystals were looped and flash-cooled in liquid N_2 at various time points.

For translocation experiments using the BcsA-2C mutant, cellulose was extended as described above and after completion of cryo-protection, sodium tetrathionate was added to a final concentration of ~1 mM or dithiobutylamine was added to a final concentration of ~100 mM, followed by incubation for 30 or 15 min, respectively. Then, UDP/$MgCl_2$ were added as described above, and crystals were harvested at various time points.

BcsA accepts UDP-Glc as well as UDP-Gal as substrates; however, because galactose is the C4 epimer of glucose, elongation of the cellulose polymer with galactose is expected to stall after a single turn over. Thus, for translocation experiments using UDP-6-thio-Gal as substrate, polymer extension was performed as described above for other substrates with the exception that the cryo solution contained 25 mM $MgCl_2$, and the 100 mM UDP-6-thio-Gal solution contained 84 mM dithiothreitol (DTT).

Data collection. Diffraction data for wild-type BcsA–BcsB and its double cysteine mutant were collected and processed as previously described[20]. Diffraction data for UDP-6-thio-Gal were collected at 6.5 keV at NE-CAT to high redundancy. Phases were obtained by molecular replacement using a search model composed of pdb 4P02 with all ligands as well as residues 332–350 (finger helix) and 499-510 (gating loop) of BcsA omitted. All ligands except for the final 4 glucose units of the cellulose polymer were subsequently added, and the models were refined in Phenix_refine[34]. Ramachandran analyses of the product-bound, substrate-bound and pre-translocation state structures identify 95.8/3.9/0.3%, 97.6/2.4/0.0% and 96.8/3.2/0.0% residues in the preferred/allowed/outlier regions, respectively. Figures were prepared using PyMol[35] and crystallographic software is supported by SBGrid[36].

UDP-CH_2-Glc soak to generate the donor-bound state. BcsA–BcsB was crystallized in the presence of 1 mM UDP-Gal, which was added to the protein/bicelle solution before mixing with the crystallization well solution. Fully-grown crystals were cryo-protected at 24 °C as described above. The crystals were then incubated with cryo-solution containing 1 mM UDP-CH_2-Glc and 10 mM $MgCl_2$ for 20 min, harvested, and flash-cooled in liquid N_2.

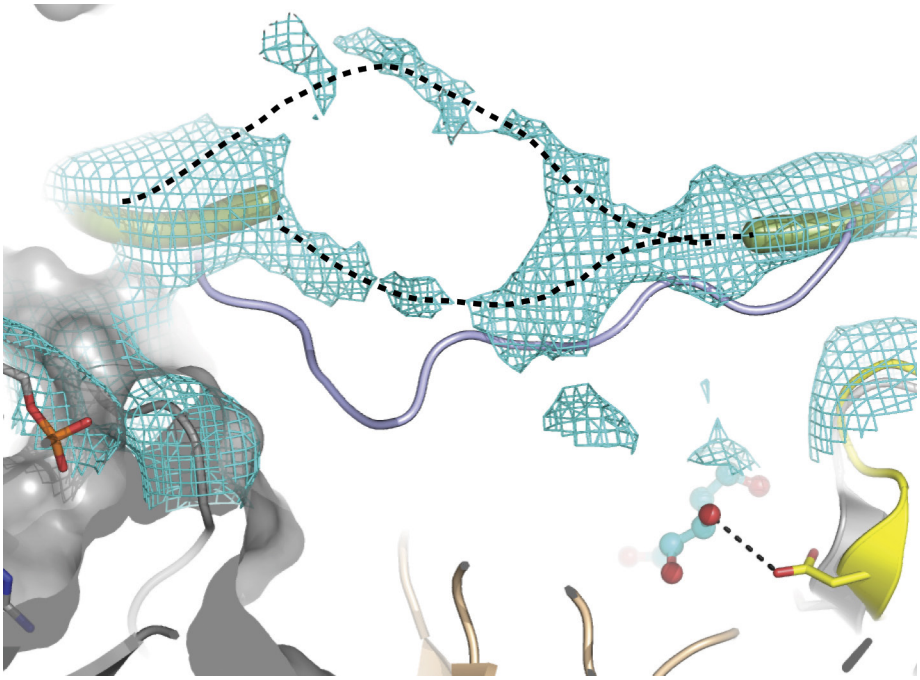
Finger helix cross-linking and activity assays. BcsA cysteine mutants were generated from the constructs described earlier[10] by using the QuikChange mutagenesis technique, and the mutant BcsA–BcsB complex was expressed and prepared as inverted membrane vesicles (IMVs) or purified and reconstituted into proteoliposomes (PLs) as previously described[7].

PLs were diluted to 125 nM (or IMVs to 12% v/v) with 125 mM NaCl and 25 mM sodium phosphate, pH 7.2, and incubated with increasing concentrations of DTT or sodium tetrathionate each for 15 min at 37 °C, before initiating cellulose biosynthesis.

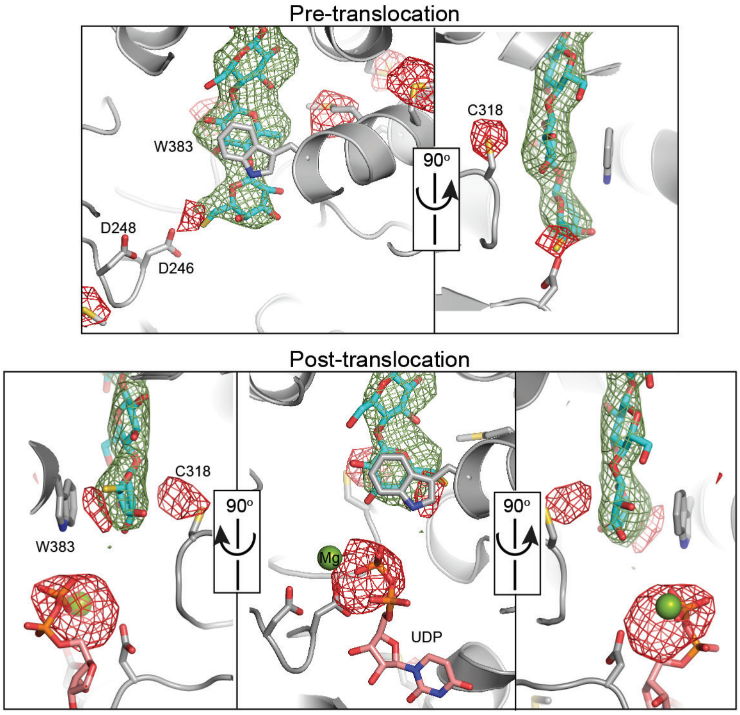
To initiate cellulose biosynthesis, c-di-GMP, UDP-Glc, UDP-[^{3}H]Glc, and $MgCl_2$ were added, giving final concentrations of 100 nM BcsA–BcsB, 20 mM sodium phosphate, pH 7.2, 100 mM NaCl, 20 mM $MgCl_2$, 5 mM UDP-Glc, 0.25 μCi UDP-[^{3}H]Glc, and 20 μM c-di-GMP. Reactions were carried out in 25 μl aliquots at 24 °C for 15 min, terminated by adding 2% SDS, and the tritium-labelled cellulose product was quantified by scintillation counting as previously described[7]. Reactions with IMVs containing BcsA Cys or I340 mutants were incubated at 24 °C for 3 h or 30 min, respectively. All experiments were performed at least in triplicate as technical replicates and error bars represent the deviations from the means.

UDP-sugar syntheses. See Supplementary Information for detailed protocols of chemical syntheses and product characterizations.

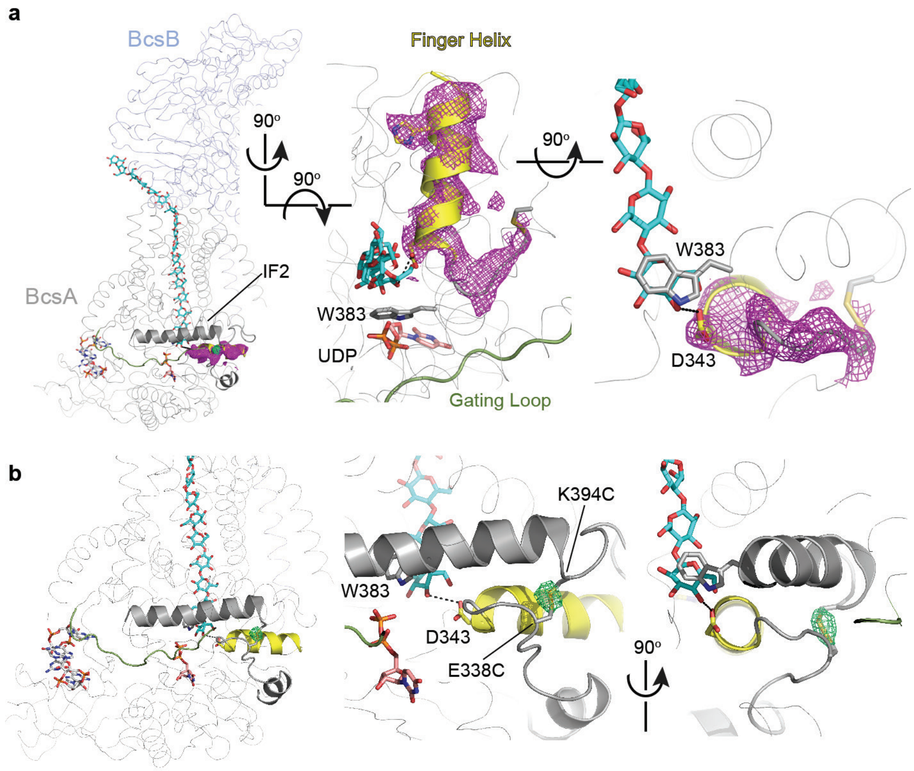
34. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
35. PyMol. *DeLano Scientific* (Sancarlos CA, USA).
36. Morin, A. *et al.* Collaboration gets the most out of software. *eLife* **2,** e01456 (2013).



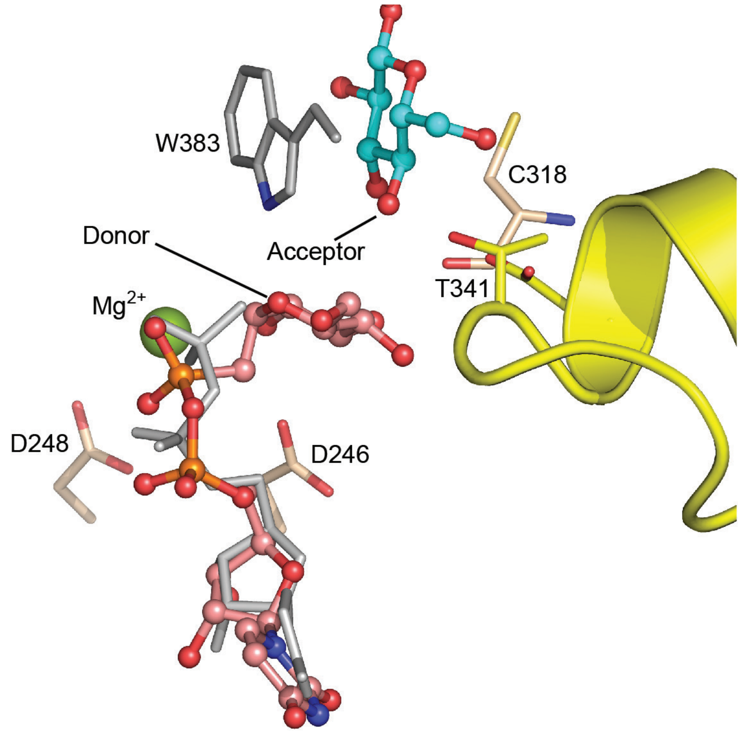
Extended Data Figure 1 | Conformational flexibility of the gating loop after cellulose extension. Unbiased Sigma-A weighted $F_o - F_c$ difference electron density of the gating loop in the pre-translocation state contoured at 2σ. The ordered part of the gating loop is shown as a thick green ribbon and two alternative backbone positions are indicated by a black dashed line. The position of the gating loop in the inserted state in the presence of a UDP molecule as observed in PDB entry 4P00 is shown as a cartoon representation coloured blue.



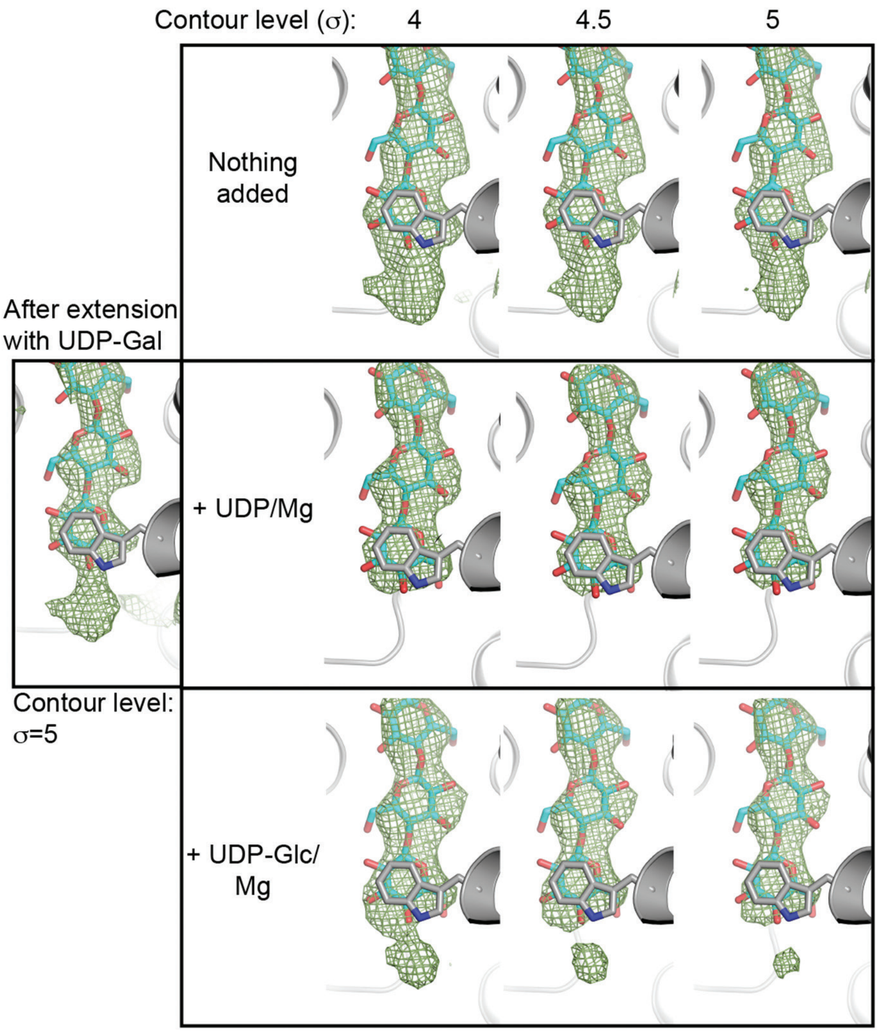
Extended Data Figure 2 | *In crystallo* translocation of a 6-thio-galactose-containing cellulose polymer. The position of the 6-thio-galactose group at the polymer's non-reducing end was determined after polymer extension (upper panel) and upon subsequent incubation with UDP/Mg^{2+} (lower panel) in an anomalous difference Fourier electron density (DANO) map. DANO peaks detected at a wavelength of 1.74 Å are shown as a red mesh contoured at 3.5σ. Unbiased Sigma-A weighted $F_o - F_c$ difference electron density for the cellulose polymer is shown as a green mesh contoured at 4σ. The cellulose polymer was extended and translocated as described in Fig. 1 with the exception that UDP-6-thio-galactose was used as substrate and Mg^{2+} was included during the initial soaking step. The extended DANO peak around Cys318 in the post-translocation state might arise from overlapping peaks originating from Cys318 and the thio-Gal unit in an opposite orientation. All Cys and Met residues close to BcsA's active site are shown as sticks. UDP is shown as sticks in violet for its carbon atoms.



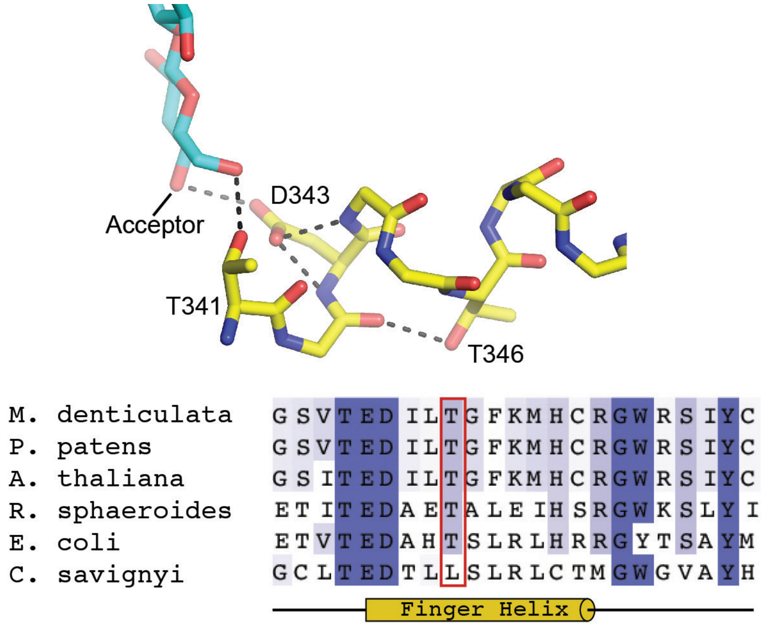
Extended Data Figure 3 | Position of the disulfide-tethered finger helix. The BcsA-2C–BcsB complex was crystallized as described for wild-type BcsA–BcsB. **a**, Unbiased Sigma-A weighted $F_o - F_c$ difference electron density of BcsA's finger helix contoured at 4σ (magenta mesh). Cellulose and BcsA's Trp383 at the entrance to the transmembrane channel are shown as sticks in cyan and grey for their carbon atoms, respectively. The finger helix and IF2 are shown as cartoon helices coloured yellow and grey, respectively. **b**, The finger helix-tethered BcsA–BcsB complex was refined in a resolution range from 34 to 3.2Å to a final R/R_{free} of 19.9/23.9% in Phenix_refine[34] with Ala residues at positions 338 and 394 of BcsA. A strong difference electron density peak indicates the position of the omitted disulfide bond in a Sigma-A weighted $F_o - F_c$ difference electron density map, (green mesh, contoured at 2σ).



Extended Data Figure 4 | Comparison of the UDP conformation in the substrate and UDP-bound states of BcsA. The substrate-bound BcsA structure was superimposed with PDB entry 4P00 by secondary structure matching in Coot. The substrate is shown as 'balls and sticks' in violet for the carbon atoms and the UDP molecule from PDB entry 4P00 is shown as grey sticks. BcsA's finger helix is shown as a yellow cartoon and the cellulose polymer is shown as cyan and red 'balls and sticks' as observed in PDB entry 4P00. Magnesium is shown as a green sphere.



Extended Data Figure 5 | UDP-Glc induced polymer translocation. The nascent cellulose polymer was extended with a chain-terminating galactose residue upon soaking BcsA–BcsB crystals with UDP-Gal. Following dilution of the substrate as described in Fig. 1, crystals were incubated for 150 min either in the absence of a nucleotide or in the presence of UDP/Mg^{2+} or UDP-Glc/Mg^{2+}, respectively. The unbiased SigmaA-weighted $F_o - F_c$ difference electron density of the nascent polymer (green mesh) is shown at three different contour levels, indicating that UDP-Glc also induces polymer translocation.



Extended Data Figure 6 | Stabilization of BcsA's finger helix by conserved residues. Top panel: stick representation of BcsA's finger helix and nascent cellulose polymer shown in yellow and cyan for their carbon atoms. The finger helix is shown as a poly-glycine helix except for the labelled residues. Bottom panel: The finger helix's 'TEDxxT' motif is conserved among pro- and eukaryotic cellulose synthases. Finger helix sequences are aligned for *Micrasterias denticulata* CesA, *Physcomitrella patens* CesA5, *Arabidopsis thaliana* CesA8, *Rhodobacter sphaeroides* and *Escherichia coli* BcsA, and *Ciona savignyi* CesA. The conserved threonine following the TED motif is indicated with a red box. Of note, the threonine residue is absent from the *Ciona* CesA sequence; however, this protein contains a serine residue at the following position, which could perform a similar function.

Extended Data Table 1 | Crystallographic data collection and refinement statistics

	Product-bound	Substrate-bound	Pre-translocation
Data collection			
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions			
a, b, c (Å)	67.3, 216.8, 221.1	68.0 216.8 220.8	67.4 218.2 220.8
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	39.43–2.95 (3.01–2.94)*	35.05–2.90 (2.96–2.90)	29.48-3.4 (3.52-3.4)
R_{pim}	0.060 (0.512)	0.094 (0.514)	0.056 (0.638)
$CC_{1/2}$^	0.995 (0.636)	0.986 (0.577)	0.997 (0.787)
Mean $I / \sigma I$	10.3 (1.6)	5.2 (1.2)	9.6 (1.3)
Completeness (%)	98.6 (82.4)	99.3 (90.0)	99.7 (100.0)
Redundancy	5.0 (4.2)	10.4 (9.1)	6.6 (6.9)
Refinement			
Resolution (Å)	34.35–2.94	34.92–2.95	29.48-3.4
No. reflections			
Total	68,776	70,259	86,075
R_{free}	3,429	3,329	4,367
R_{work} / R_{free} (%)	20.6/23.4	20.7/24.2	22.78/26.8
No. atoms			
Protein	10,673	10,725	10,618
β-1,4 glucan	198	188	199
c-di-GMP	92	92	92
UDP	25	-	-
UDP-CH_2-Glc	-	36	-
Mg^{2+}	2	2	
Lipids	89	90	69
B-factors			
Chain A	82.8	87.41	155.0
Chain B	72.2	79.4	147.0
Chain D	91.3	106.9	211.0
β-1,4 glucan	80.4	87.7	164.3
c-di-GMP	68.8	71.4	144.7
UDP	97.3		
UDP-CH_2-Glc		79.4	
Lipids	85.6	122.4	156.0
R.m.s deviations			
Bond lengths (Å)	0.003	0.003	0.004
Bond angles (°)	0.861	0.905	1.052

* Values in parentheses refer to the highest-resolution shell.

^ Correlation between intensities from random half-data sets.

ARTICLE

doi:10.1038/nature17188

Crystal structures of the M1 and M4 muscarinic acetylcholine receptors

David M. Thal[1*], Bingfa Sun[2*], Dan Feng[2], Vindhya Nawaratne[1], Katie Leach[1], Christian C. Felder[3], Mark G. Bures[4], David A. Evans[5], William I. Weis[6,7], Priti Bachhawat[2], Tong Sun Kobilka[2], Patrick M. Sexton[1], Brian K. Kobilka[2,6] & Arthur Christopoulos[1]

Muscarinic M1–M5 acetylcholine receptors are G-protein-coupled receptors that regulate many vital functions of the central and peripheral nervous systems. In particular, the M1 and M4 receptor subtypes have emerged as attractive drug targets for treatments of neurological disorders, such as Alzheimer's disease and schizophrenia, but the high conservation of the acetylcholine-binding pocket has spurred current research into targeting allosteric sites on these receptors. Here we report the crystal structures of the M1 and M4 muscarinic receptors bound to the inverse agonist, tiotropium. Comparison of these structures with each other, as well as with the previously reported M2 and M3 receptor structures, reveals differences in the orthosteric and allosteric binding sites that contribute to a role in drug selectivity at this important receptor family. We also report identification of a cluster of residues that form a network linking the orthosteric and allosteric sites of the M4 receptor, which provides new insight into how allosteric modulation may be transmitted between the two spatially distinct domains.

The M1–M5 muscarinic acetylcholine receptors constitute an important family of class A G-protein-coupled receptor (GPCRs) activated by the neurotransmitter, acetylcholine[1]. Both the M1 and M4 receptors have been associated with learning, memory, and cognition[2,3] and have emerged as attractive targets for the treatment of various central nervous system disorders, including Alzheimer's disease, schizophrenia, and drug addiction[4–6]. However, the orthosteric acetylcholine-binding site is highly conserved, and the clinical translation of compounds targeting these receptor subtypes has remained largely unsuccessful owing to adverse side effects from off-target activity at peripheral M2 and M3 receptor subtypes[7–9]. Encouragingly, muscarinic receptors possess spatially distinct allosteric binding sites that offer greater potential for selective receptor targeting[10–12], and the M1 and M4 receptors are prime examples where highly selective positive allosteric modulators (PAMs) with central nervous system activity and preclinical efficacy have been identified[4,13–17].

So far, however, the structural basis of drug action at these receptor types has been largely restricted to mutational analyses[18–21], with the only reported muscarinic receptor crystal structures being of the M2 and M3 subtypes[22,23]. Thus, to better understand the molecular basis for orthosteric and allosteric drug interactions with the M1 and M4 receptors, we sought to obtain high-resolution X-ray crystal structures of both subtypes. To gain additional insight into potential mechanisms of allosteric modulation, we complemented our findings with active-state homology modelling to rationalize the effects of targeted mutations on the interaction between a well-characterized PAM and acetylcholine at the M4 receptor.

Crystallization of the M1 and M4 receptors

To determine the structures of the M1 and M4 muscarinic receptors, we used protein engineering and lipidic cubic phase methodology[24,25]. Both receptors were crystallized in the presence of the high-affinity and clinically used inverse agonist, tiotropium (Spiriva), to stabilize the inactive state. Intracellular loop 3 (ICL3) of the M1 receptor was replaced with a T4 lysozyme fusion protein, and in the case of the M4 receptor a minimal T4 lysozyme (mT4L)[26] fusion was used to aid crystallization (Extended Data Fig. 1). It was also necessary to remove the first 21 residues of the amino (N) terminus from the M4 receptor to improve diffraction. The M1 receptor was also crystallized with the N2Q and N12Q mutations to remove glycosylation sites, and, unintentionally, an N110Q[3.37] mutation. Importantly, the binding affinities of [^{3}H]QNB (M1 receptor), [^{3}H]NMS (M4 receptor), acetylcholine, or tiotropium were not significantly different at either fusion construct compared with the wild-type receptor, suggesting that the alterations did not perturb the orthosteric site; the M1 N110Q[3.37] mutation also had no significant effect on receptor functionality in the absence of T4 lysozyme (Supplementary Table 1). The M1 and M4 structures were subsequently determined to a resolution of 2.7 Å and 2.6 Å, respectively (Extended Data Table 1).

Comparison of muscarinic receptor structures

Overall, the structures of the M1 and M4 receptors are similar to the previously solved inactive M2 and M3 receptors[22,23], with similar positioning of the seven-transmembrane (TM1–7) bundle and root mean squared deviations of 0.6–0.9 Å (Fig. 1a). Subtle differences between the receptors are observed on the extracellular and intracellular sides (Fig. 1b, c) corresponding to regions that are least conserved across the muscarinic subtypes (Extended Data Fig. 2). For example, the M2 receptor differs from the other receptors in the tilt and position of TM1 and TM7 (Fig. 1a, b). Notably, the M1 receptor was co-crystallized with a Flag peptide co-bound on the intracellular side, which makes extensive contacts with TM6 and ICL3 (Extended Data Fig. 3a, b), and probably contributes to observed differences in TM5, TM6, and a variable linkage between TM7–helix8 (Extended Data Fig. 3c). The M1-N110Q[3.37] mutation has little effect on the M1 structure other than creating a slight bulge in TM4 due to the loss of a hydrogen bond with

[1]Drug Discovery Biology and Department of Pharmacology, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, 3052, Victoria, Australia. [2]ConfometRx, 3070 Kenneth Street, Santa Clara, California 95054, USA. [3]Neuroscience, Eli Lilly, Indianapolis, Indiana 46285, USA. [4]Computational Chemistry and Chemoinformatics, Eli Lilly, Indianapolis, Indiana 46285, USA. [5]Computational Chemistry and Chemoinformatics, Eli Lilly, Sunninghill Road, Windlesham GU20 6PH, UK. [6]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA. [7]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305, USA.
*These authors contributed equally to this work.

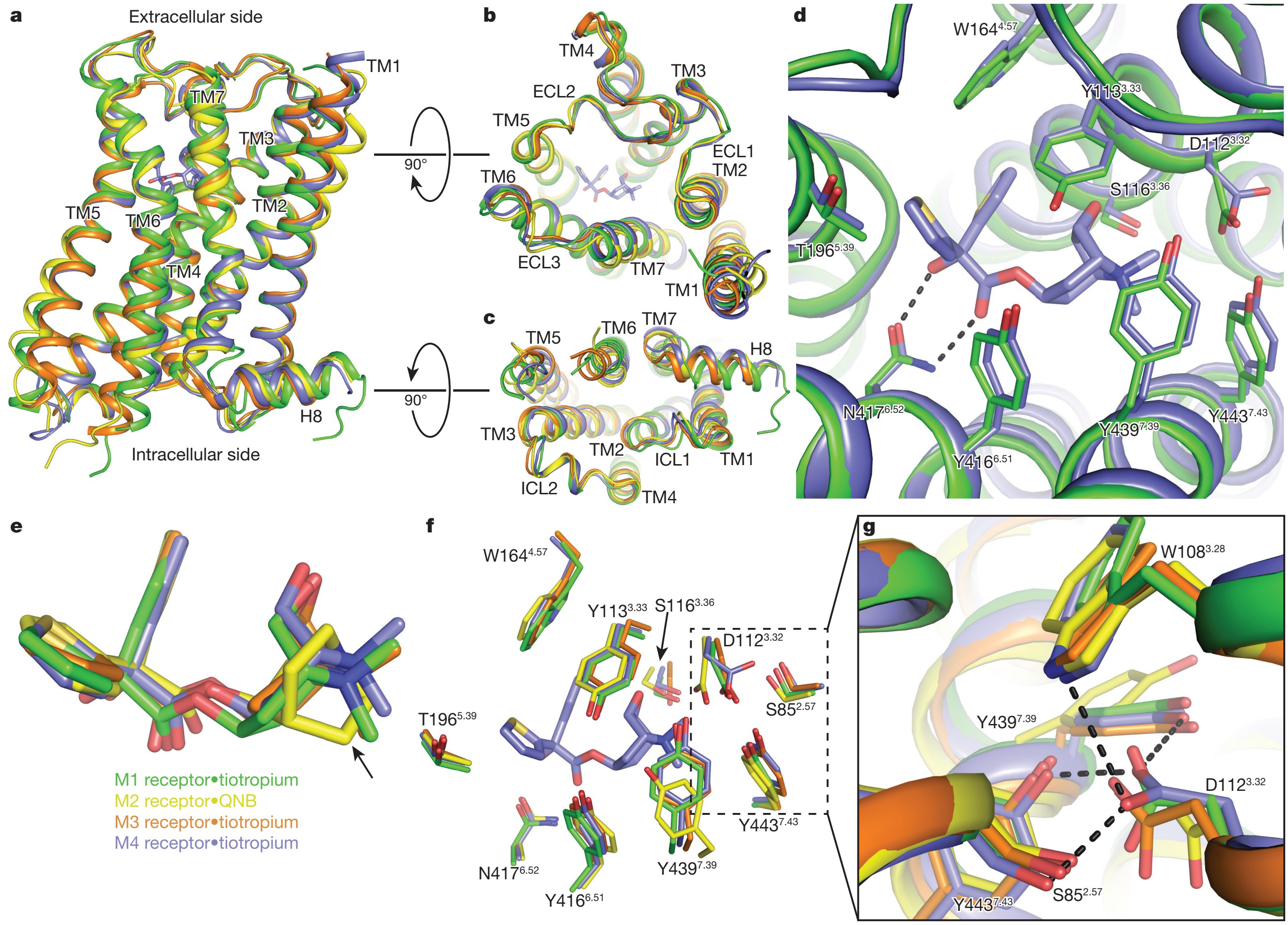


Figure 1 | Structural comparison of the M1–M4 receptors. a, The overall view of the muscarinic structures is shown as cartoons aligned to the M3 receptor, with the M1 coloured in green, M2 in yellow (PDB accession number 3UON), M3 (PDB accession number 4U15, chain A) in orange, and M4 (chain A) in blue. Root mean squared deviations for the alignment (excluding T4L fusions) of M1, M2, and M4 versus the M3 receptor are 0.86 Å, 0.81 Å, and 0.62 Å, respectively. The ligand, tiotropium, for the M4 receptor is shown as sticks and coloured according to element: carbon, light blue; oxygen, red; nitrogen, dark blue; sulfur, yellow. **b**, **c**, Comparison of the views from (**b**) the extracellular side and (**c**) the intracellular side. **d**, M1 and M4 residues involved in tiotropium binding are shown as sticks (several residues are omitted for clarity). The black dashed line indicates a bidentate hydrogen bond between N$^{6.52}$ and tiotropium. **e**, Superposition of tiotropium from the M1, M3, and M4 structures and QNB from the M2 structure. The arrow indicates the main structural difference between tiotropium and QNB. **f**, Comparison of the orthosteric binding site of the M1–M4 receptors with orthosteric site residues shown as sticks. **g**, The rotameric change of D112$^{3.32}$ is stabilized by a network of hydrogen bonds.

S$^{4.53}$ (Extended Data Fig. 3d). More interestingly, the M4 receptor was crystallized with an intact ionic lock (Extended Data Fig. 3e) a feature uncommonly seen in other GPCRs and not present in the other muscarinic structures. It is important to note that the observed differences in the intra- and extracellular sides of the receptor occur in regions that are solvent accessible or are involved in crystal packing interactions, which could contribute to the observed perturbations between subtypes; however, none of the crystal packing interactions grossly affect the structure or the core of the receptor.

Like the inactive M3 receptor, the M1 and M4 receptors were crystallized in complex with the inverse agonist, tiotropium, and this binding site is buried deep within the transmembrane core (Fig. 1d). The binding pose of tiotropium and surrounding residues between these three structures is nearly identical (Fig. 1d–f), which is not surprising given the near absolute conservation of residues lining the orthosteric site in the muscarinic family (Extended Data Fig. 2). However, this high degree of sequence conservation does not preclude the possibility of differences in tertiary structure with respect to the orthosteric site. Indeed, one surprising difference is a change in the rotamer of D112$^{3.32}$ of the M4 receptor (Fig. 1 f, g); a residue that is conserved throughout the biogenic amine GPCRs and serves as the counter ion for positively charged neurotransmitters[27]. This rotameric change points D112$^{3.32}$ away from tiotropium and is accompanied by slight movements of Y439$^{7.39}$ and Y443$^{7.43}$, allowing them to form a network of hydrogen bond interactions between D112$^{3.32}$ and S85$^{2.57}$, W108$^{3.28}$, Y439$^{7.39}$, and Y443$^{7.43}$, which is distinct from the M1, M2, and M3 muscarinic receptor structures (Fig. 1g).

Further comparison of the M1, M3, and M4 tiotropium-bound structures with the M2 receptor, which was crystallized with the structurally similar inverse agonist, QNB, also revealed considerable differences around residues D$^{3.32}$, Y$^{7.39}$, and Y$^{7.43}$. These three residues surround the amine group, which is slightly more bulky for QNB than tiotropium (Fig. 1e–g). Indeed, previous mutagenesis studies[28] on the M1 receptor revealed ligand-specific changes in binding affinities of NMS and QNB upon mutation of Y$^{7.39}$ and Y$^{7.43}$ to alanine. For the ligand NMS, which has a structurally similar tropane ring to tiotropium, a 25- and 48-fold loss of binding affinity was observed for the Y$^{7.39}$ and Y$^{7.43}$ mutations, respectively, whereas little effect was observed for QNB. This suggests a potential role for these two residues in stabilizing different inactive-state conformations with QNB potentially making compensatory

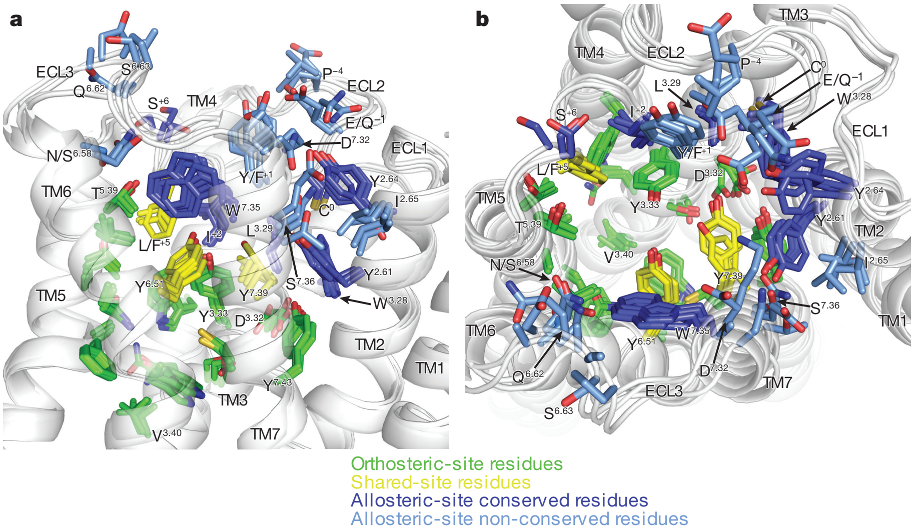


Figure 2 | Comparison of orthosteric and allosteric binding site residues across the M1–M4 receptors. a, b, The M1–M4 receptors are aligned to the M3 muscarinic receptor (PDB accession number 4U15, chain A) and are shown as grey coloured cartoons with views from the (**a**) membrane and (**b**) extracellular side. Residues in ECL2 are numbered relative to the position of the disulfide-bonded cysteine. Carbon atoms are coloured by site with the orthosteric residues in green, dark blue for allosteric conserved residues, light blue for allosteric non-conserved residues, and yellow for residues that contribute to both sites. Oxygen atoms are coloured red, nitrogen blue, and sulfur yellow. Non-conserved residues are labelled according to either the most common residue or by the residue at the M4 receptor. Residues $K^{6.62}$ (M1), $D^{6.63}$ (M1), and $S^{6.63}$ (M3) are shown as alanine owing to a lack of electron density on the side chains.

interactions unavailable to NMS. The fact that the orthosteric site of the M4 receptor is in some ways closer to the M1 than the M2 subtypes may also allow some rationalization of the relative subtype selectivity for canonical orthosteric antagonists such as pirenzepine, which has long been known to have a rank order potency of M1 > M4 > M3 > M2 (ref. 29). We performed induced fit docking (IFD) experiments of the antagonist into the inactive-state structures of the M1–M4 receptors. The overall poses for pirenzepine were very similar, with slight variability in the positioning of the methylpiperazine moiety (Extended Data Fig. 4a). However, there were still distinct differences in the orientation of residues $D^{3.32}$ and $Y^{7.39}$ between the M1, M3, and M4 subtypes versus the M2 receptor, with $D^{3.32}$ oriented towards and $Y^{7.39}$ away from the methylpiperazine group in the M2 IFD (Extended Data Fig. 4b). These differences should be interpreted with caution, as it is possible they reflect a restricted sampling in the IFD protocol, and may not be reflective of a genuine M2 specific preference. Nevertheless, these results suggest that the differences in positions of $D^{3.32}$ and $Y^{7.39}$, which surround the methylpiperazine group, could contribute to the marked difference in potency for pirenzepine between the M1 and M2 subtypes[29].

Allosteric binding and cooperativity

A comparison of all four solved muscarinic receptor structures illustrates the strikingly high degree of conservation of the residues constituting the orthosteric site (Fig. 2, green), thus providing a structural basis for the difficulty in achieving subtype selectivity when targeting this region. In contrast, muscarinic receptors possess a large extracellular vestibule that contains residues contributing to an allosteric site. As shown in Fig. 2 (blue), comparison of these residues reveals a striking divergence between subtypes, owing to differences in amino-acid composition (Extended Data Fig. 2) and likely additional tertiary structure changes that arise as a consequence of the dynamic nature of the extracellular loop regions. Also shown in Fig. 2 (yellow) are residues that have been previously suggested to form the 'roof' of the orthosteric site and 'floor' of the allosteric site[20,30]. These 'shared' residues show an intermediate degree of tertiary structure divergence between subtypes compared with the orthosteric and allosteric site residues, and are conserved among all five subtypes with the exception of the M2 receptor where L in ECL2 is replaced by F.

Comparison of the electrostatic surface potential of each receptor (Fig. 3) also reveals distinct differences in both the shape and charge distribution of the allosteric site and can explain why some of the best-studied muscarinic receptor allosteric modulators are cationic compounds[31]. For example gallamine[32], a prototypical negative allosteric modulator of muscarinic receptors, has a binding potency order of M2 > M1, M4 > M3, M5 (ref. 33). The acidic EDGE sequence (Fig. 3b) of the M2 receptor has been shown to be important for gallamine affinity and cooperativity; indeed, replacement of M1 residues LAGQ with the EDGE (Fig. 3b) significantly improved gallamine affinity at the M1 receptor[33].

Interestingly, inspection of our M4 receptor data also revealed that the precipitant, polyethylene glycol 300 (PEG 300), is able to occupy the allosteric binding site of the inactive-state receptor (Extended Data Fig. 5), a finding consistent with the recent structure of the M3-mT4L receptor[26]. Surrounding the PEG 300 molecule are residues that form the allosteric site from the top regions of TM2, TM3, and TMs 5–7 (Extended Data Fig. 5b). Furthermore, PEG 300 sits immediately above the aromatic cage composed of $Y113^{3.33}$, $Y416^{6.51}$, $Y439^{7.39}$, and $Y443^{7.43}$ (Extended Data Fig. 5a). These residues have been implicated in regulating the dissociation of antagonists from the orthosteric binding site[34], and we confirmed the ability of PEG 300 to act as an allosteric modulator in its own right through its ability to retard the dissociation of [^{3}H]NMS in a concentration-dependent manner with a calculated apparent affinity of approximately 10 mM for the [^{3}H]NMS-occupied M4 receptor (Extended Data Fig. 5c, d).

Our finding above illustrates an inherent difficulty in obtaining inactive-state structures with prototypical negative allosteric modulators bound in the open muscarinic extracellular vestibule, as PEG 300 is a required precipitant and is present at concentrations of over 1.0 M. However, a recent breakthrough was the solution of the active-state structure of the related M2 muscarinic receptor bound to a high efficacy agonist, iperoxo, in the absence or presence of the PAM, LY2119620, which preferentially bound in a more tightly closed vestibule that arises in the active-state[35]. Because the M4 receptor is most closely related to the M2 subtype, and M4 receptor PAMs are highly pursued as novel therapeutic agents[4,36], we undertook a combined mutagenesis and molecular modelling study to complement our structural work and gain additional insights into mechanisms governing positive allosteric modulation at this muscarinic receptor subtype.

We investigated the interaction between the well-characterized PAM, LY2033298 (refs 13–15, 20, 21), and the cognate agonist, acetylcholine.

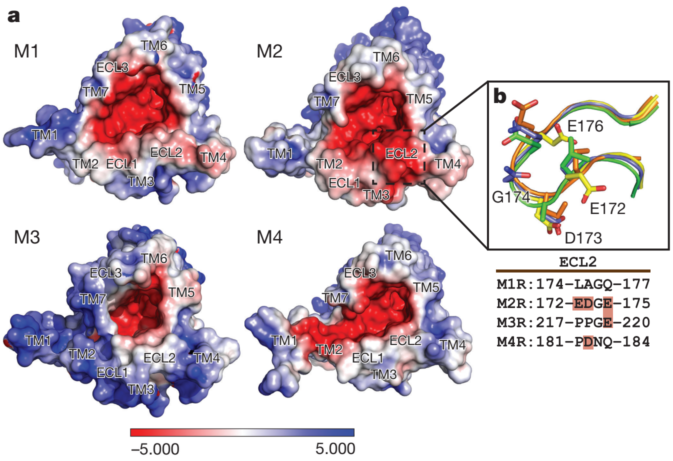


Figure 3 | Electrostatic and surface properties of the different muscarinic receptor structures. a, Electrostatic potentials ($+5kT/e$ in blue and $-5kT/e$ per electron in red) mapped on the surfaces of the M1–M4 receptor structures calculated at pH 7.0 using the programs PDB2PQR[43,44] and APBS[45]. **b**, Residues in ECL2 that make up the EDGE sequence at the M2 receptor and the corresponding regions at the other subtypes are shown as sticks. Negatively charged residues in the sequence alignment are coloured red.

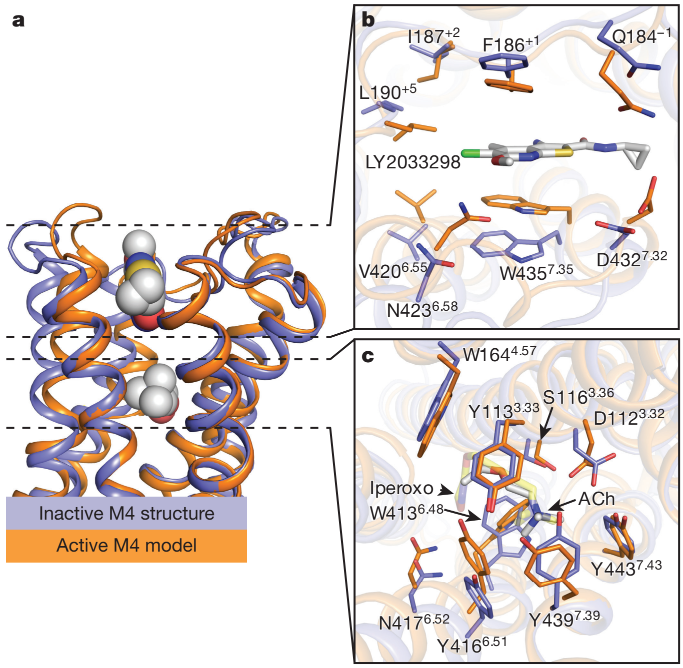


Figure 4 | Model of an active M4 receptor bound to acetylcholine and LY2033298. a, Comparison of the M4•tiotropium receptor structure (blue) versus an active-state model (orange) bound to acetylcholine and LY2033298 on the basis of the active M2•iperoxo•LY2119620 structure (PDB accession number 4MQT) as viewed from the membrane. The active M4 model was aligned to the M4•tiotropium structure (chain A, excluding T4L) with a root mean squared deviation of 0.9 Å. **b**, **c**, Cross-sectional views of the (**b**) allosteric site and (**c**) orthosteric site as viewed from the extracellular side with a 90° rotation relative to the membrane from **a**. Residues surrounding each site are shown as sticks (several removed for clarity). Acetylcholine and LY2033298 are shown as (**a**) spheres or (**b**, **c**) sticks and coloured according to element: carbon, cyan; oxygen, red; nitrogen, blue; sulfur, yellow; chlorine, green. Acetylcholine is in the *trans* conformation and aligns in a similar pose to iperoxo (**c**, transparent yellow sticks).

On the basis of the recent structure of the active M2 receptor bound to the LY2033298 congener, LY2119620 (refs 35, 37, 38), it is likely that such PAMs bind to an essentially pre-formed closed state of the extracellular vestibule. As such, residues whose mutation might alter the cooperativity between acetylcholine and LY2033298 fall into three general categories: (1) those that make tighter contacts with the ligands in the closed state than the open state; (2) those that are immobilized by the binding of either ligand, such that the entropic cost is paid by the first binding event; (3) non-ligand-contact residues that alter the free energy of activation of the receptor and thus the open to closed transition. We chose to focus on residues within and between the extracellular vestibule and orthosteric sites, which are likely to reflect the first two categories; mutagenesis of non-contact residues that govern the free energy of receptor transitions are beyond the scope of the current work.

Because prior mutagenesis studies suggested a role for aromatic residues in receptor interaction with LY2033298, we generated alanine mutations of selected aromatic residues near the top of the receptor and applied an allosteric ternary complex model to the data (Methods) to determine the effect of each mutation on the affinity of acetylcholine (K_A) or LY2033298 (K_B) for the free receptor and the magnitude of positive cooperativity (α) between the two ligands. We also chose to investigate selected (non-aromatic) residues that line the proximal and distal ends of ECL2, given the important role this region plays in the binding of modulators to the extracellular vestibule[18,39–41]. The results of these experiments are summarized in Supplementary Tables 2–4 and include prior mutagenesis results from our laboratory for the same set of ligands. To rationalize our findings, we used the recent active state M2 receptor structure as a template to generate a homology model of the M4 receptor bound to acetylcholine and LY2033298, and compared this with our inactive state crystal structure (Fig. 4 and Extended Data Fig. 6).

The most dramatic effect on the affinity of the PAM was noted upon mutation of W435$^{7.35}$ at the top of TM7, with a complete loss in LY2033298 binding, similar to our previous observations[21] upon alanine substitution of F186$^{\mathrm{ECL2}}$ (Fig. 5, Extended Data Fig. 7 and Supplementary Tables 3 and 4). Alanine mutations of residues Y113$^{3.33}$, Y416$^{6.51}$, and Y439$^{7.39}$, which form the roof of the orthosteric site, led to significant decreases in cooperativity. A slight increase in modulator affinity and significant decrease in cooperativity was also noted with mutation of Y89$^{2.61}$, together with our prior identification of residues W108$^{3.28}$ and L109$^{3.29}$ as likely contributors to the PAM binding pocket[20]. Comparison of our inactive state structure to the active state model now provides a mechanistic rationale for our findings, specifically a contraction of the extracellular vestibule that results predominantly in an inward movement of N423$^{6.58}$, F186$^{\mathrm{ECL2}}$, and W435$^{7.35}$ allowing π-stacking interactions to occur with the modulator in the active state (Fig. 4b). For the acetylcholine-binding pocket, there is a contraction of the pocket mediated by an inward movement of the top of TM6 to accommodate the large difference in size between acetylcholine and tiotropium resulting in significant movement of residues Y416$^{6.51}$, N417$^{6.52}$, W413$^{6.48}$, and Y439$^{7.39}$ (Fig. 4c). Additionally, D112$^{3.32}$ is reoriented to interact with the choline head-group of acetylcholine, and is no longer stabilized by the same hydrogen bond network that is seen in the inactive state (Fig. 1g).

Importantly, mapping of the amino-acid residues that significantly affect the cooperativity between acetylcholine and LY2033298 upon mutation also identified, for the first time, a network that appears to link the allosteric and orthosteric sites, involving the interface between TMs 2, 3, 6, and 7, and extending along the top of ECL2 (Fig. 5; orange coloured residues); this network is consistent with views of allosteric modulation that propose a preferred energetic link between orthosteric and allosteric sites[42] but, to our knowledge, has never been directly mapped before in a GPCR. Interestingly, a comparison of the side-chain locations between the inactive M4 structure and active M4 model for residues in the allosteric network reveals that the majority of residues at the TM2/3/7 interface that contribute to cooperativity are not predicted to undergo appreciable movement between states, whereas comparison of residues further away from the interface (F186$^{\mathrm{ECL2}}$, Y416$^{6.51}$,

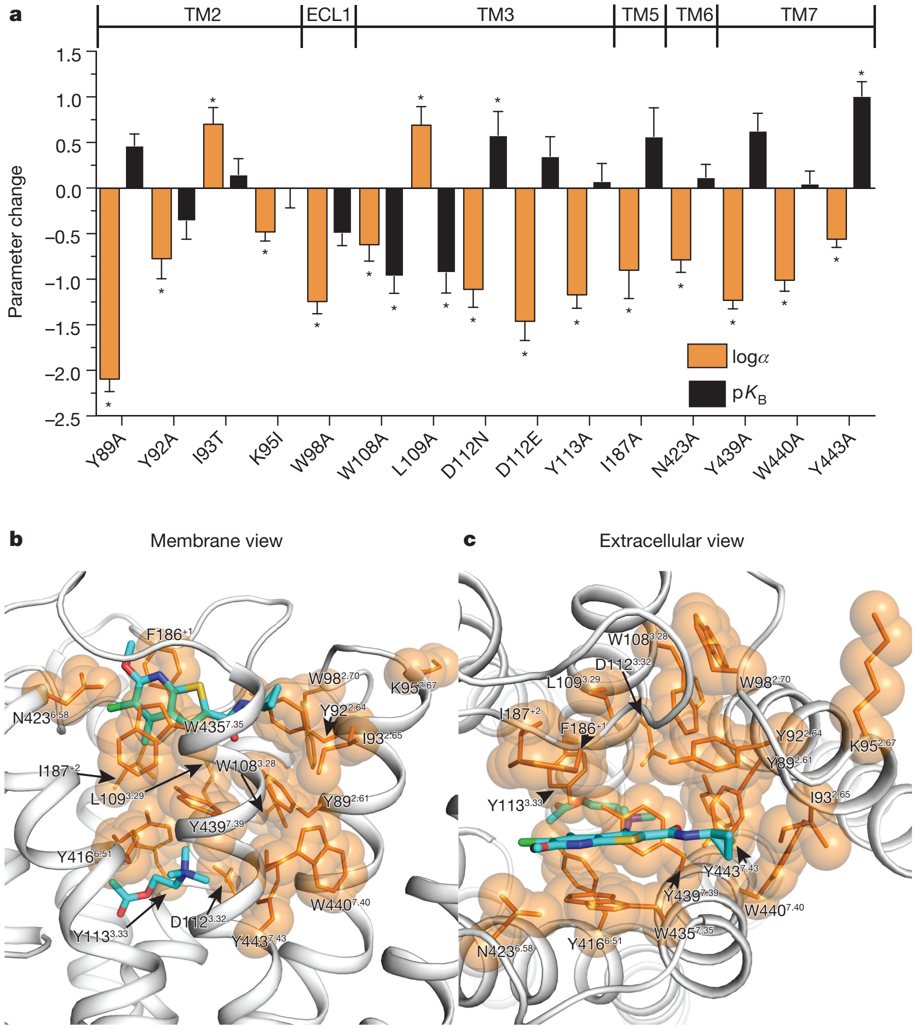


Figure 5 | A cooperativity network at the M4 receptor. a, Changes in either LY2033298 binding affinity (ΔpK_B, coloured black) or cooperativity ($\Delta\log\alpha$, coloured orange) relative to wild-type M4 are shown for each mutation. Data represent the mean ± s.e.m. from at least three experiments performed in duplicate. Statistical differences between pharmacological parameters at wild-type versus mutant M4 receptors are indicated by asterisks and were determined by one-way analysis of variance with Dunnett's post hoc test, where $P < 0.01$ (Supplementary Table 3) or $P < 0.05$ (for previously determined mutations; Supplementary Table 4) were considered statistically significant. Cooperativity and binding values for $F186^{+1}$ and $W435^{7.35}$ were not determined owing to a lack of LY2033298 binding (see Supplementary Tables 2–4). **b**, **c**, Residues from **a** were mapped onto the M4 active-state model and coloured as orange sticks with translucent spheres with views from (**b**) the membrane and (**c**) the extracellular side. LY2033298 and acetylcholine are shown as sticks and coloured the same as in Fig. 4.

$N423^{6.58}$, and $W435^{7.35}$) are predicted to move significantly between the two states (Extended Data Fig. 8). The TM2/3/7 interface, which forms part of the hydrophobic core of the receptor, may act as a hinge mediating conformational rearrangements in the extracellular vestibule between the inactive (open extracellular vestibule) and active (closed extracellular vestibule) states of the receptor. Disruption of this hinge by mutagenesis alters the packing interactions within the interface and might change the energetic barrier between the open and closed conformations of the receptor leading to either an increase or decrease in PAM cooperativity. Thus, binding of a PAM to the allosteric site might stabilize the conformation of the allosteric network residues that are otherwise found in a more dynamic state. Presumably, structures of the inactive state and active M4 model described here represent the lowest energy conformations, as they were obtained using crystallography, or are based on the X-ray structures of the active M2 receptor[35] (Extended Data Fig. 8).

Another noteworthy feature of LY2033298 is that it is selective towards the M4 receptor versus the M1 receptor when tested against acetylcholine[15]. This difference in selectivity could arise either through differential binding affinities of LY2033298 or through a difference in the cooperativity between LY2033298 and acetylcholine between the two subtypes (Extended Data Fig. 9).

Conclusions

Muscarinic receptors remain important drug targets, and designing molecules to selectively target the orthosteric binding site has proved challenging, as highlighted by the lack of prominent differences between the receptor subtypes. Alongside the previously determined M2 and M3 structures, the M1 and M4 structures presented here now offer a near complete view of the inactive state of this important subfamily of GPCRs. Excitingly, comparison of these structures clearly reveals a divergence in residues lining the allosteric site, highlighting the importance of this region for designing selective drugs. Moreover, our enriched structure–function analysis of the M4 receptor indicates that it is possible to combine crystal structure and mutagenesis data to uncover new insights into GPCR allosteric modulation, and our results point to the TM2/3/7 interface as a network for further studies on the mechanistic basis of allostery at class A GPCRs. Together with the recent solution of the inactive M2 and M3 receptors, as well as the active and PAM-bound M2 receptor, our study has contributed to an emerging picture of mechanisms of allostery at a therapeutically important receptor family that may facilitate the design of novel agents targeting a variety of CNS disorders while avoiding peripheral off-target effects.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 June 2015; accepted 29 January 2016.
Published online 9 March 2016.

1. Wess, J., Eglen, R. M. & Gautam, D. Muscarinic acetylcholine receptors: mutant mice provide new insights for drug development. *Nature Rev. Drug Discov*. **6,** 721–733 (2007).
2. Hasselmo, M. E. The role of acetylcholine in learning and memory. *Curr. Opin. Neurobiol.* **16,** 710–715 (2006).
3. Hasselmo, M. E. & Giocomo, L. M. Cholinergic modulation of cortical function. *J. Mol. Neurosci.* **30,** 133–135 (2006).
4. Kruse, A. C. *et al.* Muscarinic acetylcholine receptors: novel opportunities for drug development. *Nature Rev. Drug Discov*. **13,** 549–560 (2014).
5. Kruse, A. C., Hu, J., Kobilka, B. K. & Wess, J. Muscarinic acetylcholine receptor X-ray structures: potential implications for drug development. *Curr. Opin. Pharmacol.* **16,** 24–30 (2014).
6. Foster, D. J., Jones, C. K. & Conn, P. J. Emerging approaches for treatment of schizophrenia: modulation of cholinergic signaling. *Discov. Med.* **14,** 413–420 (2012).
7. Shekhar, A. *et al.* Selective muscarinic receptor agonist xanomeline as a novel treatment approach for schizophrenia. *Am. J. Psychiatry* **165,** 1033–1039 (2008).
8. Bodick, N. C. *et al.* Effects of xanomeline, a selective muscarinic receptor agonist, on cognitive function and behavioral symptoms in Alzheimer disease. *Arch. Neurol.* **54,** 465–473 (1997).
9. Heinrich, J. N. *et al.* Pharmacological comparison of muscarinic ligands: historical versus more recent muscarinic M1-preferring receptor agonists. *Eur. J. Pharmacol.* **605,** 53–56 (2009).
10. Conn, P. J., Christopoulos, A. & Lindsley, C. W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nature Rev. Drug Discov.* **8,** 41–54 (2009).
11. Digby, G. J., Shirey, J. K. & Conn, P. J. Allosteric activators of muscarinic receptors as novel approaches for treatment of CNS disorders. *Mol. Biosyst.* **6,** 1345–1354 (2010).
12. Keov, P., Sexton, P. M. & Christopoulos, A. Allosteric modulation of G protein-coupled receptors: a pharmacological perspective. *Neuropharmacology* **60,** 24–35 (2011).
13. Suratman, S. *et al.* Impact of species variability and 'probe-dependence' on the detection and *in vivo* validation of allosteric modulation at the M4 muscarinic acetylcholine receptor. *Br. J. Pharmacol.* **162,** 1659–1670 (2011).
14. Leach, K. *et al.* Molecular mechanisms of action and *in vivo* validation of an M4 muscarinic acetylcholine receptor allosteric modulator with potential antipsychotic properties. *Neuropsychopharmacology* **35,** 855–869 (2010).
15. Chan, W. Y. *et al.* Allosteric modulation of the muscarinic M4 receptor as an approach to treating schizophrenia. *Proc. Natl Acad. Sci. USA* **105,** 10978–10983 (2008).
16. Shirey, J. K. *et al.* A selective allosteric potentiator of the M1 muscarinic acetylcholine receptor increases activity of medial prefrontal cortical neurons and restores impairments in reversal learning. *J. Neurosci.* **29,** 14271–14286 (2009).
17. Ma, L. *et al.* Selective activation of the M1 muscarinic acetylcholine receptor achieved by allosteric potentiation. *Proc. Natl Acad. Sci. USA* **106,** 15950–15955 (2009).
18. Abdul-Ridha, A. *et al.* Molecular determinants of allosteric modulation at the M1 muscarinic acetylcholine receptor. *J. Biol. Chem.* **289,** 6067–6079 (2014).
19. Abdul-Ridha, A. *et al.* Mechanistic insights into allosteric structure-function relationships at the M1 muscarinic acetylcholine receptor. *J. Biol. Chem.* **289,** 33701–33711 (2014).
20. Leach, K., Davey, A. E., Felder, C. C., Sexton, P. M. & Christopoulos, A. The role of transmembrane domain 3 in the actions of orthosteric, allosteric, and atypical agonists of the M4 muscarinic acetylcholine receptor. *Mol. Pharmacol.* **79,** 855–865 (2011).
21. Nawaratne, V., Leach, K., Felder, C. C., Sexton, P. M. & Christopoulos, A. Structural determinants of allosteric agonism and modulation at the M4 muscarinic acetylcholine receptor: identification of ligand-specific and global activation mechanisms. *J. Biol. Chem.* **285,** 19012–19021 (2010).
22. Kruse, A. C. *et al.* Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482,** 552–556 (2012).
23. Haga, K. *et al.* Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **482,** 547–551 (2012).
24. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4,** 706–731 (2009).
25. Chun, E. *et al.* Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* **20,** 967–976 (2012).
26. Thorsen, T. S., Matt, R., Weis, W. I. & Kobilka, B. K. Modified T4 lysozyme fusion proteins facilitate G protein-coupled receptor crystallogenesis. *Structure* **22,** 1657–1664 (2014).
27. van Rhee, A. M. & Jacobson, K. A. Molecular architecture of G protein-coupled receptors. *Drug Dev. Res.* **37,** 1–38 (1996).
28. Lu, Z.-L., Saldanha, J. W. & Hulme, E. C. Transmembrane domains 4 and 7 of the M_1 muscarinic acetylcholine receptor are critical for ligand binding and the receptor activation switch. *J. Biol. Chem.* **276,** 34098–34104 (2001).
29. Caulfield, M. P. & Birdsall, N. J. M. International Union of Pharmacology. XVII. Classification of muscarinic acetylcholine receptors. *Pharmacol. Rev.* **50,** 279–290 (1998).
30. Goodwin, J. A., Hulme, E. C., Langmead, C. J. & Tehan, B. G. Roof and floor of the muscarinic binding pocket: variations in the binding modes of orthosteric ligands. *Mol. Pharmacol.* **72,** 1484–1496 (2007).
31. Gregory, K. J., Sexton, P. M. & Christopoulos, A. Allosteric modulation of muscarinic acetylcholine receptors. *Curr. Neuropharmacol.* **5,** 157–167 (2007).
32. Stockton, J. M., Birdsall, N. J., Burgen, A. S. & Hulme, E. C. Modification of the binding properties of muscarinic receptors by gallamine. *Mol. Pharmacol.* **23,** 551–557 (1983).
33. Gnagey, A. L., Seidenberg, M. & Ellis, J. Site-directed mutagenesis reveals two epitopes involved in the subtype selectivity of the allosteric interactions of gallamine at muscarinic acetylcholine receptors. *Mol. Pharmacol.* **56,** 1245–1253 (1999).
34. Tautermann, C. S. *et al.* Molecular basis for the long duration of action and kinetic selectivity of tiotropium for the muscarinic M3 receptor. *J. Med. Chem.* **56,** 8746–8756 (2013).
35. Kruse, A. C. *et al.* Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* **504,** 101–106 (2013).
36. Conn, P. J., Jones, C. K. & Lindsley, C. W. Subtype-selective allosteric modulators of muscarinic receptors for the treatment of CNS disorders. *Trends Pharmacol. Sci.* **30,** 148–155 (2009).
37. Schober, D. A., Croy, C. H., Xiao, H., Christopoulos, A. & Felder, C. C. Development of a radioligand, [^{3}H]LY2119620, to probe the human M_2 and M_4 muscarinic receptor allosteric binding sites. *Mol. Pharmacol.* **86,** 116–123 (2014).
38. Croy, C. H. *et al.* Characterization of the novel positive allosteric modulator, LY2119620, at the muscarinic M_2 and M_4 receptors. *Mol. Pharmacol.* **86,** 106–115 (2014).
39. Khoury, E., Clément, S. & Laporte, S. A. Allosteric and biased g protein-coupled receptor signaling regulation: potentials for new therapeutics. *Front. Endocrinol.* **5,** 68 (2014).
40. Scarselli, M., Li, B., Kim, S. K. & Wess, J. Multiple residues in the second extracellular loop are critical for M_3 muscarinic acetylcholine receptor activation. *J. Biol. Chem.* **282,** 7385–7396 (2007).
41. Avlani, V. A. *et al.* Critical role for the second extracellular loop in the binding of both orthosteric and allosteric G protein-coupled receptor ligands. *J. Biol. Chem.* **282,** 25677–25686 (2007).
42. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286,** 295–299 (1999).
43. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32,** W665–W667 (2004).
44. Dolinsky, T. J. *et al.* PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35,** W522–W525 (2007).
45. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98,** 10037–10041 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Lopez for generating initial M4 homology models. This work was funded by Program Grant APP1055134 of the National Health and Medical Research Council (NHMRC) of Australia (A.C., P.M.S.). Portions of this work were supported by a Lilly Research Award Program grant. W.I.W. and B.K.K. were supported by the Mathers Foundation. A.C. is a Senior Principal, and P.M.S. a Principal, Research Fellow of the NHMRC. GM/CA @ APS has been funded in whole or in part with federal funds from the National Cancer Institute (Y1-CO-1020) and the National Institute of General Medical Science (Y1-GM-1104). Use of the Advanced Photon Source was supported by the US Department of Energy, Basic Energy Sciences, Office of Science, under contract number W-31-109-ENG-38.

Author Contributions D.M.T. performed cloning, protein expression, purification, crystallization, data collection, structure refinement, and radioligand binding assays on the M4 receptor. D.F. purified and crystallized the M1 receptor. B.S. performed data collection and structure refinement on the M1 receptor. K.L., V.N., and D.M.T. performed mutagenesis and radioligand binding studies that examined the effects of amino-acid substitutions on ligand pharmacology. C.C.F., M.G.B., and D.E. provided the pirenzepine IFD and active-state M4 homology model. P.B. generated the active-state model of the M1 receptor. T.S.K. supervised the M1 muscarinic receptor production and purification. W.I.W. supervised structure refinement. B.K.K., P.M.S., and A.C. provided overall project supervision. D.M.T. and A.C. wrote the manuscript.

Author Information Atomic coordinates and structure factors for the M1 and M4 receptors, respectively, have been deposited in the Protein Data Bank (PDB) under accession numbers 5CXV and 5DSG. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.M.S. (patrick.sexton@monash.edu), B.K.K. (kobilka@stanford.edu) or A.C. (arthur.christopoulos@monash.edu).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

M1 and M4 receptor expression and purification. The human M4 muscarinic receptor gene (http://www.cdna.org) was cloned into a modified pFastBac1 vector to give a receptor containing an N-terminal Flag epitope tag and a carboxy (C)-terminal 8× histidine tag. Residues 226–389 of ICL3 were removed and replaced by a minimal Cys-free T4 lysozyme fusion protein[26]. The human M1 muscarinic receptor gene was also cloned into the modified pFastBac1 vector, and residues 219–354 of ICL3 were removed and replaced by a Cys-free T4 lysozyme fusion protein. Both fusion proteins were expressed using the Bac-to-Bac Baculovirus Expression System (Invitrogen) in *Sf9* cells. Cells were infected at a density of 4.0×10^6 to 5.0×10^6 cells per millilitre, treated with 10 μM atropine, and harvested at 60 h. Receptor was solubilized and purified in the presence of tiotropium as previously described for the M3 (ref. 22) receptor using Ni-NTA chromatography, Flag affinity chromatography, and size-exclusion chromatography. The N terminus of the M4 receptor was removed by cleavage with HRV 3C protease at a concentration of 2% (w/w) during concentration of the receptor before size-exclusion chromatography (~2 h at 4 °C). After size-exclusion chromatography, purified receptor was concentrated to 85 absorbance units (~50 mg ml^{-1}) and flash frozen in small aliquots using liquid nitrogen.

Pharmacology of crystallization constructs. Sf9 cells expressing wild-type M4 or M4-mT4L receptor, as described above, were pelleted and washed with PBS three times for 1 h each to remove any bound atropine. Cells were resuspended in binding buffer (10 mM HEPES pH 7.5, 100 mM NaCl, and 10 mM $MgCl_2$) and flash frozen with liquid nitrogen. Saturation binding assays were performed using approximately 20,000 cells per well with 9 different concentrations of [^{3}H]NMS in a total volume of 0.5 ml for 3 h at 37 °C. Competition binding assays with acetylcholine and tiotropium were performed in the presence of a fixed concentration of [^{3}H]NMS over 10 different concentrations of ligand for 3 h at 37 °C. Non-specific binding was measured in the presence of 10 μM atropine, and reactions were harvested by rapid filtration through GF/B filters. Data were analysed using Prism 6.0d. Similar methods were applied for binding assays using wild type M1 and M1–T4L, except that [^{3}H]QNB was used as the radioligand.

Crystallization. Purified M1-T4L•tiotropium and M4-mT4L•tiotropium were crystallized using lipid cubic phase technology. Each receptor was reconstituted by mixing the protein solution into 10:1 (w/w) monoolein:cholesterol (Sigma) in 1:1.5 parts w/w protein:lipid ratio using the two-syringe method[24]. For the M1 receptor, samples of 50 nl (20–40 nl for M4) were spotted onto 96-well glass plates and overlaid with 800 nl (600 nl for M4) of precipitant solution for each well using a Gryphon LCP (Art Robbins Instruments). Glass plates were then sealed using a glass cover film and incubated at 20 °C. Initial crystals for the M1 receptor formed after 24 h in conditions containing 33% PEG 300, 100 mM sodium acetate, and 100 mM Bis-Tris Propane (pH 8.0). For the M4 receptor, initial crystals formed after 24 h in conditions containing 25–40% PEG 300, 50–100 mM EDTA (pH 8.0), and 100 mM MES (pH 5.5–6.5). M1 and M4 crystals were harvested using mesh grid loops (MiTeGen) and stored in liquid nitrogen before use.

Data collection, processing, and structure determination. X-ray diffraction data were collected at the Advanced Photon Source at Argonne National Laboratories at GM/CA beamline 23ID-D. Crystals were located by initial rastering using an 80 μm by 30 μm beam with fivefold attenuation and 1 s exposure. Regions that contained strong diffraction were then sub-rastered using a 10 μm collimated beam with fivefold attenuation. Data were then collected with the 10 μm beam using no attenuation with 1–2 s exposures and 1 degree oscillations. To prevent radiation damage, data were collected in wedges of 3–10° before moving onto either a different site on the same crystal or a new crystal. Diffraction data were processed using HKL2000 (M1 receptor) or XDS[46] (M4 receptor) and statistics are summarized in Extended Data Table 1. Both structures were solved by molecular replacement using Phaser[47]. For the M1 receptor, the inactive M3 structure[22] (PDB accession number 4DAJ) was split into its receptor and T4L components and used as corresponding search models. The refinement was performed using Refmac5 (ref. 48) with manual building in Coot[49]. For the M4 receptor, the inactive M2 structure[23] (PDB accession number 3UON) and the inactive M3-mT4L[26] (PDB accession number 4U15) were used as search models for the receptor and mT4L fusion domains, respectively. The resulting model was completed by iterative refinement in Phenix[50] and manual building with Coot[49]. MolProbity[51] was used for structure validation, and figures were prepared using PyMol[52]. Final refinement statistics are reported in Extended Data Table 1.

Induced fit docking of pirenzepine. The inactive state structures of M1, M2, M3 (PDB 4U15, chain B), and M4 (chain A) receptors were processed by the protein preparation wizard of the Schrodinger 2014-2 suite[53], after deleting the lysozyme insertion region. Missing side chains were added by Prime and hydrogens refined by minimization with the OPLS2.1 force field. Binding grids were defined using the default settings in Glide, centring the grid on the crystallized orthosteric ligand in each case. The PEG ligand in the extracellular vestibule of M3 and M4 receptors was deleted before grid generation. The ligand, pirenzepine, was treated with ligprep software to generate initial protonated 3D structures. Compound structures were docked using the induced fit docking protocol with default settings, which involves the use of the OPLS_2005 force field to refine residues around poses docked by Glide SP, followed by redocking into the generated receptor conformations, also with Glide SP. The poses with the lowest induced fit score were selected. This scoring function takes into account an estimate of the protein conformational penalty along with a protein–ligand interaction docking score.

Molecular modelling of active M4 receptor. A homology model of a human active-state M4 receptor was constructed using the Prime program implemented in Maestro version 2014.1 from Schrodinger. The crystal structure of the M2 receptor with an orthosteric and allosteric agonist bound (PDB accession number 4MQT) was used as a template to build the M4 model. The M2–M4 sequence alignment generated by Prime needed no adjustment owing to the overall significant sequence homology between the two isoforms. The initial M4 receptor model was built with the allosteric ligand (LY2119620) present in the M2 crystal structure bound in the M4 allosteric site and with iperoxo bound in the orthosteric site (as also present in the M2 structure). The binding mode of LY2119620 in M4 was used as a guide to manually dock LY2033298 into the M4 allosteric binding site. In addition, iperoxo from the M4 model was manually modified into acetylcholine (ACh). The M4-ACh-LY2033298 complex was then subjected to 500 steps of energy minimization (MacroModel implemented in Maestro 2014.1 from Schrödinger[53]) to optimize key interactions in the binding sites. The resulting model of ACh and LY2033298 bound to M4 was used in subsequent modelling studies described in this paper.

Molecular modelling of active M1 receptor. The active state of the M1 receptor was modelled on the basis of the active state structure of M2 bound to iperoxo (PDB accession number 4MQT), using the automated protein structure homology modelling web server Swiss-Model[54,55]. The nanobody structure was removed and the resulting coordinates were used as a template to model the M1 primary sequence without intracellular loop 3 residues (residues 213–240). The model was built using Promod-II, minimized by steepest descent energy minimization using a GROMOS96 force field and the quality was assessed by the QMEAN scoring function. ACh and LY2033298 were docked in the M1 homology model using Swiss-Dock[56], using steric and chemical considerations such as shape, charge complimentary, and keeping the protein structure constant. The top-scoring clusters were evaluated manually on the basis of chemical and steric considerations to pick the favourable pose. Owing to static docking, the top four ACh poses did not affect the docking results for LY2033298. For ACh, the selected pose is in the trans conformation similar to the M4•ACh•LY2033298 model. Finally, the structures with the ligand were energy minimized using Chimera with standard Steepest Descent and Conjugate Gradient steps.

Receptor mutagenesis and generation of cell lines. DNA encoding the human M4 mAChR with a triple HA[20] or cmyc[21] tag at its N terminus was subjected to QuikChange site-directed mutagenesis (Stratagene) to generate M4 mAChR sequences with the desired amino-acid substitutions. DNA constructs in pEF5/frt/V5 (Invitrogen) were stably expressed in Flp-In-CHO cells (Invitrogen), which were maintained in high-glucose Dulbecco's modified Eagle's medium containing 10% FBS, 16 mM HEPES, and 400 μg ml^{-1} hygromycin B. Mycoplasma testing was performed regularly on cell lines using the MycoAlertTM kit (Lonza); cell lines were mycoplasma-free before experiments were conducted.

Radioligand binding assays. Cell membranes were prepared as described previously[14,57]. [^{3}H]QNB affinity (K_A) at the M4 WT receptor and mutants was determined by saturation binding assays, performed by incubating varying concentrations of [^{3}H]QNB with 10–100 μg of membranes at 37 °C for 1 h, in a final volume of 0.5–1 ml binding buffer (20 mM HEPES, 100 mM NaCl, and 10 mM $MgCl_2$ at pH 7.4).

Radioligand inhibition binding assays were performed by co-incubating 10–100 μg of membranes with a K_A concentration of [^{3}H]QNB (determined in saturation assays, Supplementary Table 2) and varying concentrations of the non-radiolabelled test compound in 0.5–1 ml binding buffer in the presence of the guanine nucleotide, GppNHp (100 μM), which was used to promote receptor/G-protein uncoupling. These experiments determined the concentration of ACh that inhibited 20% [^{3}H]QNB binding, defined as the 20% inhibitory concentration (IC_{20}), which was used in subsequent interaction studies between [^{3}H]QNB, ACh, and LY2033298. These experiments were performed by co-incubating 10–100 μg of membranes, an IC_{20} concentration of ACh, and a K_A concentration of [^{3}H]QNB with increasing concentrations of LY2033298 in binding buffer containing GppNHp (100 μM). The reaction was left to reach equilibrium for 3 h at 37 °C. For all experiments, non-specific binding was defined in the presence of

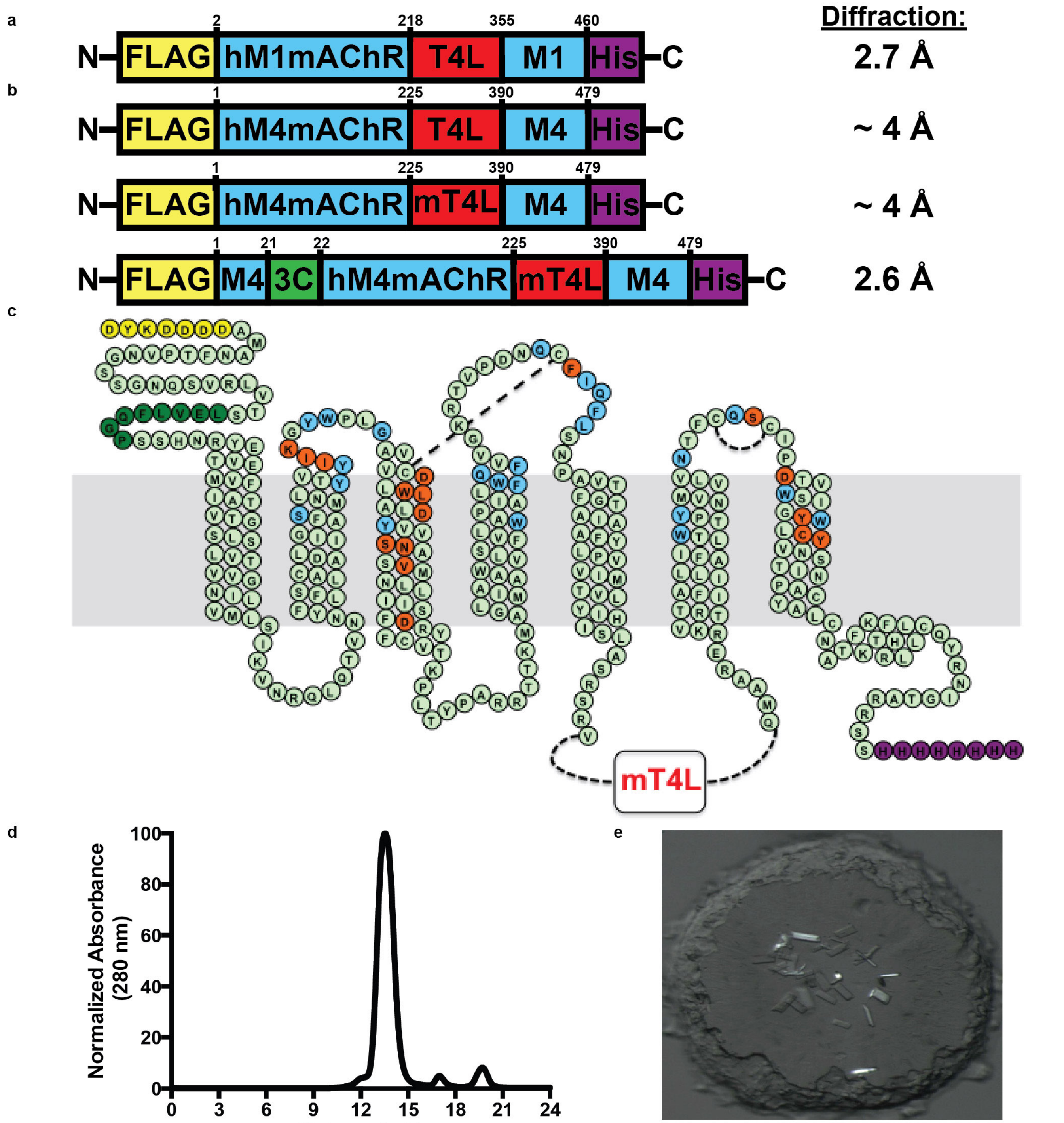
$10\,\mu$M atropine, total binding was determined in the absence of the test ligand, and vehicle effects were determined with 0.1% dimethylsulfoxide (DMSO). The assays were terminated by vacuum filtration through GF-B glass fibre filters, which were washed three times with ice-cold 0.9% NaCl. [^{3}H]QNB radioactivity was measured using a Packard 1600 TR liquid scintillation beta counter. Owing to a lack [^{3}H]QNB binding, affinity data for W164A$^{4.57}$ were determined from functional pERK1/2 experiments performed as previously described[20,21].

Data analysis. Data were analysed using Prism (GraphPad). For radioligand saturation binding, non-specific and total binding data were analysed as described previously[58]. Inhibition binding curves between [^{3}H]QNB and ACh were fitted to a one-site binding model[58]. Interaction experiments between [^{3}H]QNB, ACh, and LY2033298 were fitted to the following allosteric ternary complex model[20,21,59]:

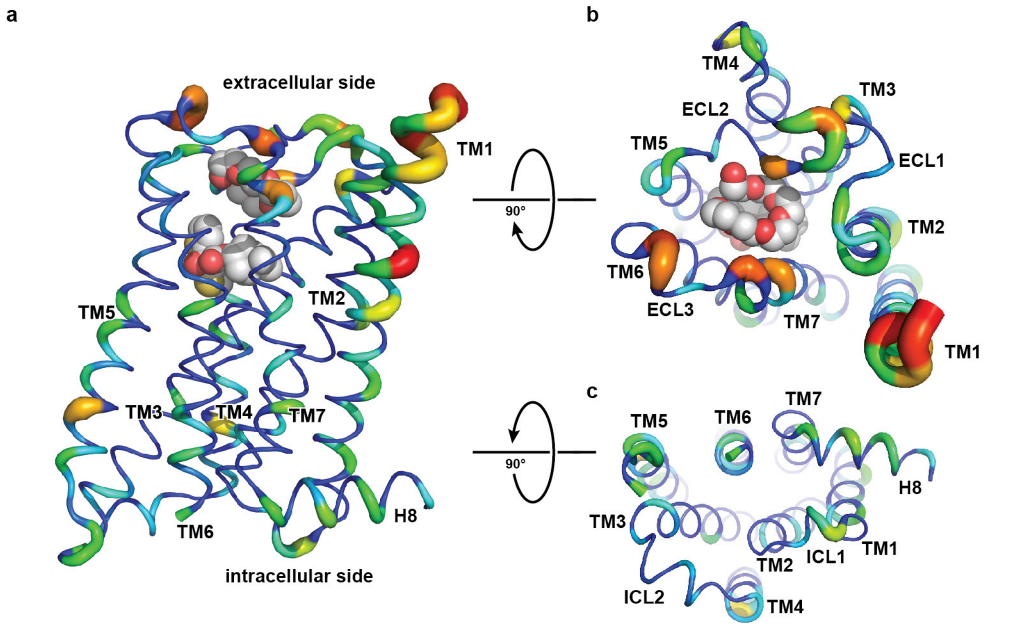
$$Y = \frac{\mathrm{B_{max}}\,[\mathrm{A}]}{[\mathrm{A}] + \left(\frac{K_\mathrm{A} K_\mathrm{B}}{\alpha'[\mathrm{B}] + K_\mathrm{B}}\right)\left(1 + \frac{[\mathrm{I}]}{K_\mathrm{I}} + \frac{[\mathrm{B}]}{K_\mathrm{B}} + \frac{\alpha[\mathrm{I}][B]}{K_\mathrm{I} K_\mathrm{B}}\right)}$$

where Y is the specific radioligand binding, B_{max} is the total number of receptors, [A], [B], and [I] are the concentrations of radioligand, allosteric modulator, and unlabelled orthosteric ligand, respectively, K_A, K_B, and K_I are the equilibrium dissociation constants of the radioligand, allosteric modulator, and unlabelled orthosteric ligand, respectively, and α' and α are the cooperativity factors between allosteric modulator and the radioligand or unlabelled orthosteric ligand, respectively. The value of α' was taken as 1 when the binding of [^{3}H]QNB changed by less than 10% at 10^{-5} M LY2033298 relative to zero LY2033298, and was fixed as such for all analyses. Otherwise, the value of α' was determined using a global fit to the allosteric ternary complex model. Statistical differences between pharmacological parameters at wild-type versus mutant M4 receptors were determined by one-way analysis of variance with Dunnett's post hoc test, where $P < 0.01$ was considered statistically significant.

46. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
47. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
48. Skubák, P., Murshudov, G. N. & Pannu, N. S. Direct incorporation of experimental phase information in model refinement. *Acta Crystallogr. D* **60,** 2196–2201 (2004).
49. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
50. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
51. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
52. Schrodinger, L. The PyMOL Molecular Graphics System, version 1.7.0.3 (2010).
53. Schrödinger release 2014-2: Maestro, version 9.7 (2014).
54. Guex, N., Peitsch, M. C. & Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis* **30** (Suppl. 1), S162–S173 (2009).
55. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22,** 195–201 (2006).
56. Grosdidier, A., Zoete, V. & Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* **39,** W270–W277 (2011).
57. Nawaratne, V. *et al.* New insights into the function of M4 muscarinic acetylcholine receptors gained using a novel allosteric modulator and a DREADD (designer receptor exclusively activated by a designer drug). *Mol. Pharmacol.* **74,** 1119–1131 (2008).
58. Motulsky, H. & Christopoulos, A. *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting* (Oxford Univ. Press, 2004).
59. Ehlert, F. J. Estimation of the affinities of allosteric ligands using radioligand binding and pharmacological null methods. *Mol. Pharmacol.* **33,** 187–194 (1988).
60. Celniker, G. *et al.* ConSurf: using evolutionary data to raise testable hypotheses about protein function. *Isr. J. Chem.* **53,** 199–206 (2013).
61. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38,** W529–W533 (2010).
62. Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33,** W299–W302 (2005).
63. Glaser, F. *et al.* ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19,** 163–164 (2003).
64. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
65. Molecular Operating Environment (MOE) (Chemical Computing Group, 2015).
66. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).



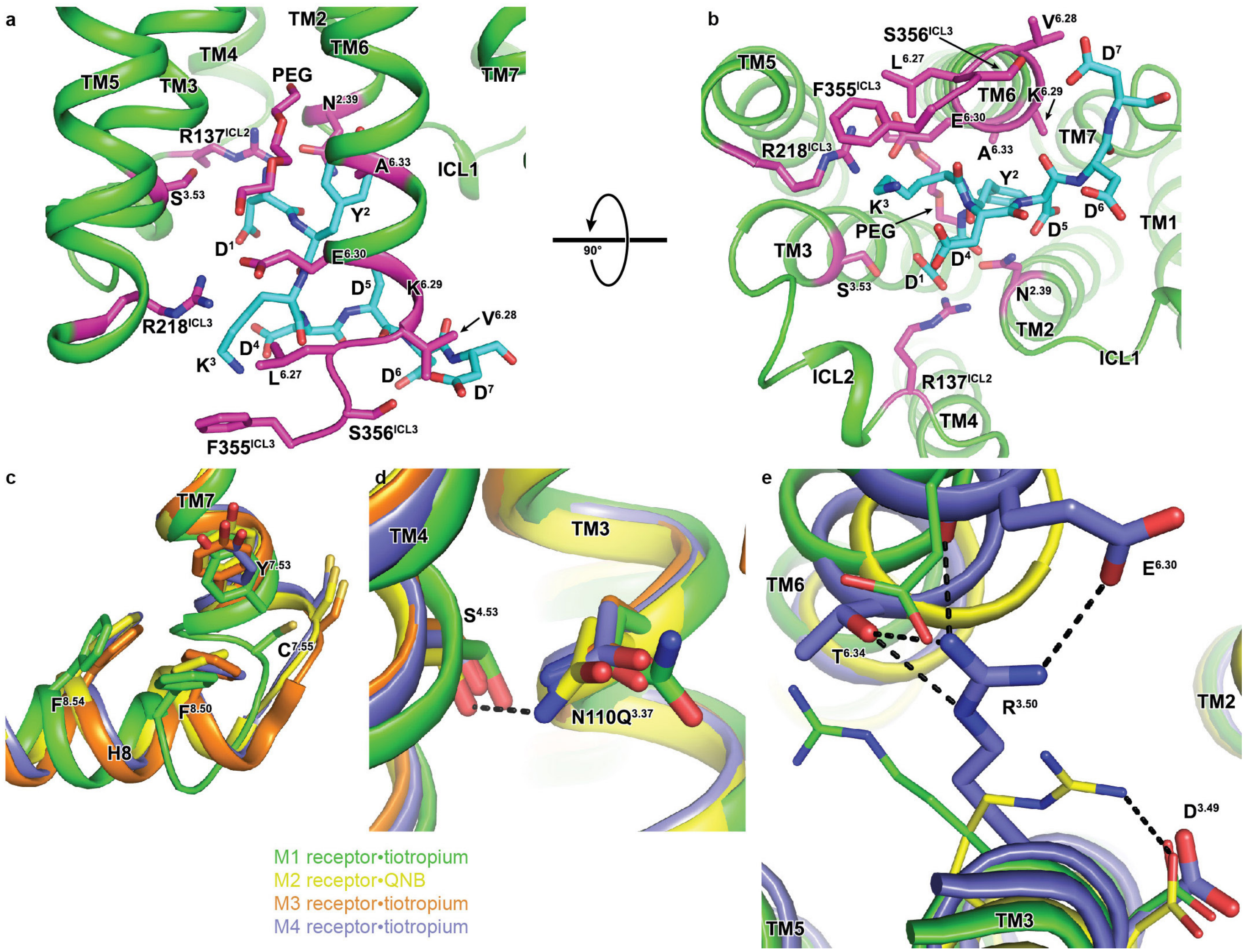
Extended Data Figure 1 | Crystallization construct design, purification and crystallization. **a**, **b**, Crystallization constructs used for the (**a**) M1 receptor and (**b**) M4 receptor. All constructs contain an N-terminal Flag epitope (yellow), C-terminal histidine tag (purple), and a T4L lysozyme fusion protein (red). For the M4 receptor, initial constructs diffracted out to 4 Å; however, the diffraction data appeared to suffer from a lattice translocation disorder and were unsolvable. The final crystallization construct contained a shortened N terminus with an HRV 3C cleavage site, shown in dark green, and a minimal T4 lysozyme fusion (mT4L)[26], shown in red. **c**, Snake-plot diagram of the best diffracting M4 mAChR construct coloured according to **a**. Residues coloured blue are single-point mutations from this study, and residues coloured orange are previously studied mutations[20,21]. **d**, Size-exclusion chromatography trace of purified monodispersed M4-mT4L bound to tiotropium. **e**, Crystals of M4-mT4L obtained in lipidic cubic phase and observed under circularly polarized light.



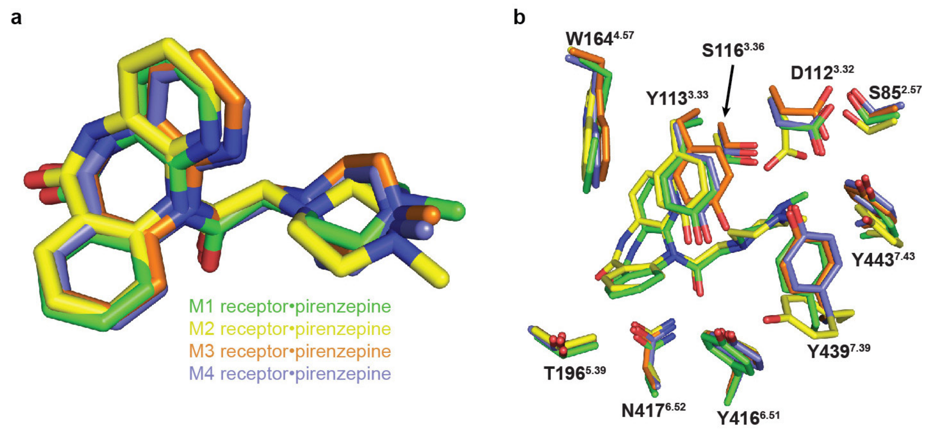
d

```
N-----------TM1------------ICL1---------TM2-------------ECL1-
M1_receptor 15-VLAPGKGPWQVAFIGITTGLLSLATVTGNLLVLISFKVNTELKTVNNYFLLSLACADLIIGTFSMNLYTTYLLMGHWALG 94
M2_receptor 13-ALTSPYKTFEVVFIVLVAGSLSLVTIIGNILVMVSIKVNRHLQTVNNYFLFSLACADLIIGVFSMNLYTLYTVIGYWPLG 92
M3_receptor 58-DPLGGHTVWQVVFIAFLTGILALVTIIGNILVIVSFKVNKQLKTVNNYFLLSLACADLIIGVISMNLFTTYIIMNRWALG 137
M4_receptor 22-SSHNRYETVEMVFIATVTGSLSLVTVVGNILVMLSIKVNRQLQTVNNYFLFSLACADLIIGAFSMNLYTVYIIKGYWPLG 101
M5_receptor 20-QPLERHRLWEVITIAAVTAVVSLITIVGNVLVMISFKVNSQLKTVNNYLLSLACADLIIGIFSMNLYTTYILMGRWALG 99
                 ::  *   :. ::* *: **:**::*:*** .*:*****:*:********** :****:* * : . *.**

-----TM3------------ICL2------TM4---------
M1_receptor TLACDLWLALDYVASNASVMNLLLISFDRYFSVTRPLSYRAKRTPRRAALMIGLAWLVSFVLWAPAILFWQYLVGERTVL 174
M2_receptor PVVCDLWLALDYVVSNASVMNLLIISFDRYFCVTKPLTYPVKRTTKMAGMMIAAAWVLSFILWAPAILFWQFIVGVRTVE 172
M3_receptor NLACDLWLAIDYVASNASVMNLLVISFDRYFSITRPLTYRAKRTTKRAGVMIGLAWVISFVLWAPAILFWQYFVGKRTVP 217
M4_receptor AVVCDLWLALDYVVSNASVMNLLIISFDRYFCVTKPLTYPARRTTKMAGLMIAAAWVLSFVLWAPAILFWQFVVGKRTVP 181
M5_receptor SLACDLWLALDYVASNASVMNLLVISFDRYFSITRPLTYRAKRTPKRAGIMIGLAWLISFILWAPAILCWQYLVGKRTVP 179
            :.******:***.*********:*******.:*:**:* .:**.: *.:**. **::**:******* **:.** ***

----ECL2---TM5-------------ICL3--------------TM6
M1_receptor AGQCYIQFLSQPIITFGTAMAAFYLPVTVMCTLYWRIYRETENRARELAALQGSETPGKGGGSSSS----FSLVKEKKAA 364
M2_receptor DGECYIQFFSNAAVTFGTAIAAFYLPVIIMTVLYWHISRASKSRIKKDKKEPVANQDPVSPSLVQG----PPPSREKKVT 386
M3_receptor PGECFIQFLSEPTITFGTAIAAFYMPVTIMTILYWRIYKETEKRTKELAGLQASGTEAETENFVHP----MSLVKEKKAA 490
M4_receptor DNQCFIQFLSNPAVTFGTAIAAFYLPVVIMTVLYIHISLASRSRVHKHRPEGPKEKKAKTLAFLKS----QMAARERKVT 399
M5_receptor LDECQIQFLSEPTITFGTAIAAFYIPVSVMTILYCRIYRETEKRTKDLADLQGSDSVTKAEKRKPA----VVLVKERKAA 441
             .:* ***:*:. :*****:****:** :*   ** :*    :..* :.                          :*:*.:

---------TM6-----------ECL3----TM7------------H8-----------C
M1_receptor RTLSAILLAFILTWTPYNIMVLVSTFCKDCVPETLWELGYWLCYVNSTINPMCYALCNKAFRDTFRLLLLCRWDKRRWRK-444
M2_receptor RTILAILLAFIITWAPYNVMVLINTFCAPCIPNTVWTIGYWLCYINSTINPACYALCNATFKKTFKHLLMCHYKNIGATR 466
M3_receptor QTLSAILLAFIITWTPYNIMVLVNTFCDSCIPKTFWNLGYWLCYINSTVNPVCYALCNKTFRTTFKMLLLCQCDKKKRRK-570
M4_receptor RTIFAILLAFILTWTPYNVMVLVNTFCQSCIPDTVWSIGYWLCYVNSTINPACYALCNATFKKTFRHLLLCQYRNIGTAR 479
M5_receptor QTLSAILLAFIITWTPYNIMVLVSTFCDKCVPVTLWHLGYWLCYVNSTVNPICYALCNRTFRKTFKMLLLCRWKKKKVEE-521
            :*: *******:**:***:***:.***  *:* *.* :******:***:** ****** :*: **: **:*:  :    .
```

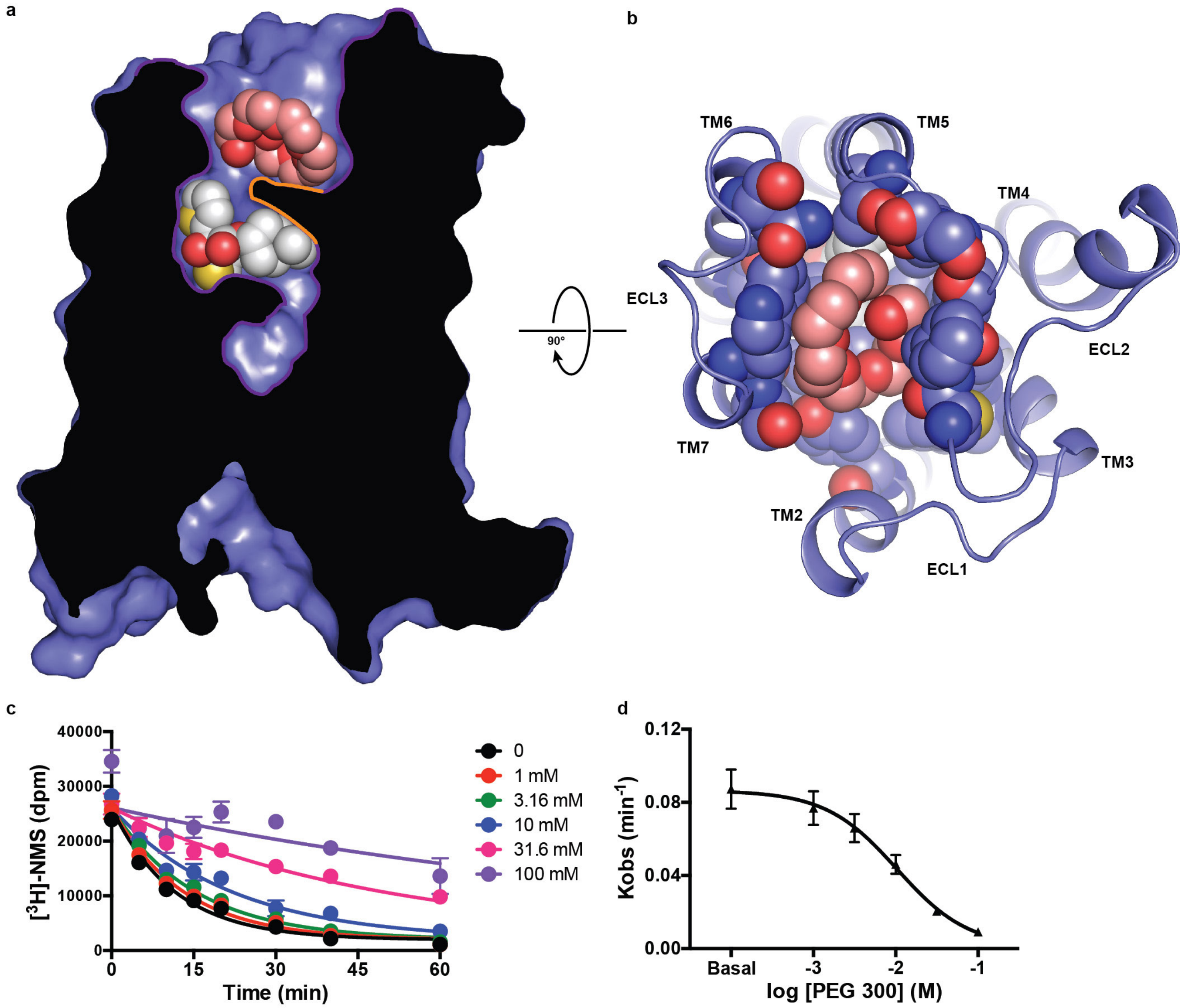
Extended Data Figure 2 | Sequence conservation across the muscarinic receptor subfamily. a–c, The sequence alignment of the human M1–M5 receptors (**d**) was determined on the ConSurf server to calculate amino-acid conservation scores[60,61]. Conservation scores for each residue were mapped[62,63] onto the M4 structure and coloured as a gradient from blue (highly conserved) to red (least conserved) with views from the (**b**) extracellular and (**c**) intracellular sides. The radius of the cartoon increases as the residues at each position become more poorly conserved. Tiotropium and PEG 300 from the M4 structure are shown as spheres and coloured with carbon in white, oxygen in red, nitrogen in blue, and sulfur in yellow. **d**, Amino-acid sequences of the human M1–M5 receptors were aligned using the ClustalW2 server[64]. Alpha helical regions are shown as blue boxes as determined by the consensus of the M1–M4 structures. The most conserved residue in each TM (X.50) is in bold lettering. Regions of the N terminus, C terminus, and ICL3 regions are removed for space and clarity. Insertion points of the T4 lysozyme fusion proteins between TM5 and TM6 are underlined with bold lettering. Residues from the orthosteric binding-site are highlighted in red and allosteric binding-site residues in blue. Residues that contribute to both sites are coloured in yellow.



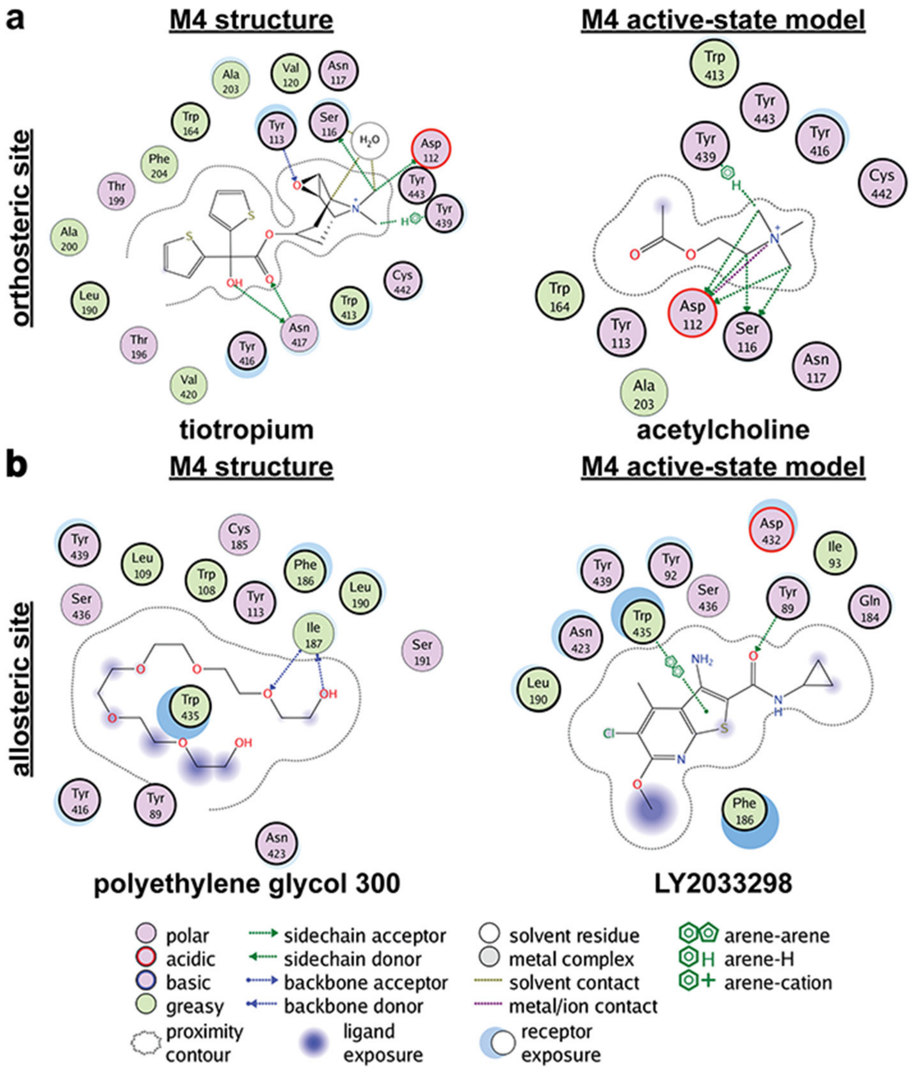
Extended Data Figure 3 | Distinct structural features for the M1 and M4 receptors. The receptors shown are aligned and coloured as in Fig. 1. **a**, **b**, The M1 receptor was crystallized with the Flag peptide (DYKDDDD; coloured cyan sticks) co-bound on the cytoplasmic surface. Residues of the M1 receptor within 4 Å of the Flag peptide are shown as magenta coloured sticks with views from the (**a**) membrane and (**b**) cytoplasmic side. **c**, The linkage between TM7 and helix 8 of the M1 receptor undergoes a bend starting with a change in rotamer of residue $Y^{7.53}$, which may be a result of perturbations in TM6 due to the Flag peptide. **d**, The M1-N110$Q^{3.37}$ mutation causes a slight bulge in TM4 due to the loss of a hydrogen bond with $S^{4.53}$. **e**, Chain B of the M4 receptor has an intact ionic lock with $R^{3.50}$ forming hydrogen bonds with $T^{6.34}$ and $E^{6.30}$.



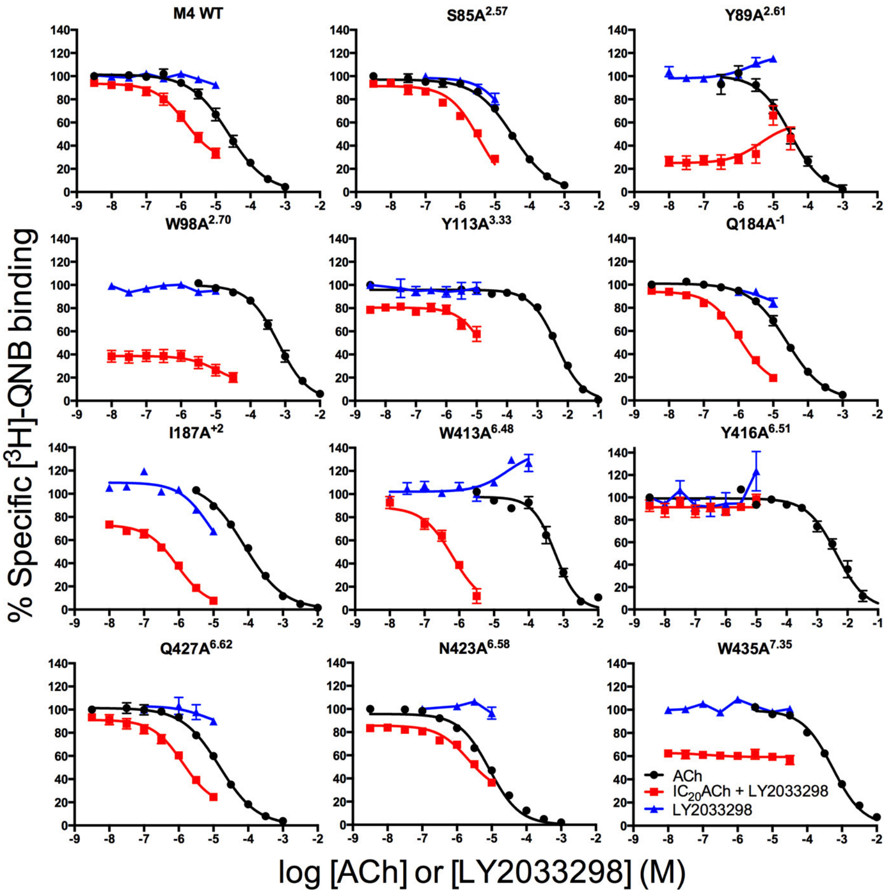
Extended Data Figure 4 | Induced fit docking of pirenzepine into the M1–M4 structures. The receptors shown are aligned and coloured as in Fig. 1. **a**, Superposition of the poses of pirenzepine from the IFD experiments. **b**, Comparison of the pirenzepine poses for the M1 and M4 receptor with residues that contribute to the orthosteric site of the M1–M4 receptors (several residues omitted for clarity).



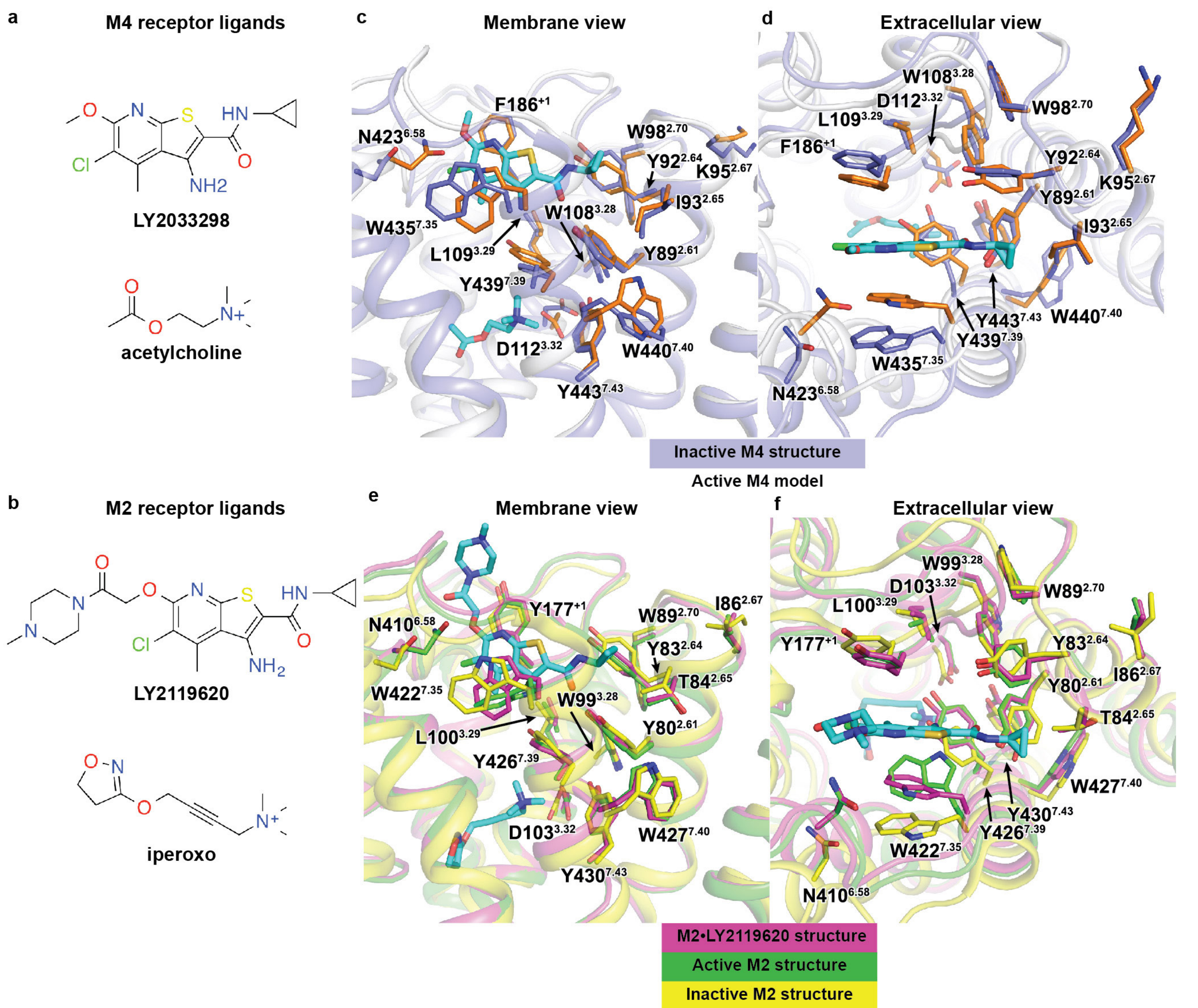
Extended Data Figure 5 | PEG 300 occupies the allosteric binding site of the inactive M4 receptor. a, The cross section of the solvent accessible surface area of the M4 receptor is coloured blue. Tiotropium and PEG 300 are shown as spheres with respective carbons coloured white and peach. The aromatic cage of covering tiotropium is highlighted in orange **b**, View from the extracellular side with residues that contact PEG 300 shown as spheres. **c**, Dissociation kinetics of [^{3}H]NMS in the presence of PEG 300. [^{3}H]NMS was incubated with M4-mT4L membranes at 37 °C for 3 h, followed by addition of 10 μM atropine ± PEG 300 at the indicated concentrations and time points. Representative data from three experiments, performed in duplicate, fitted to a one-phase exponential decay are shown. **d**, PEG 300 has an apparent binding affinity for the NMS-occupied receptor of approximately 10 mM ($\log(IC_{50}) = -1.95 \pm 0.02$).



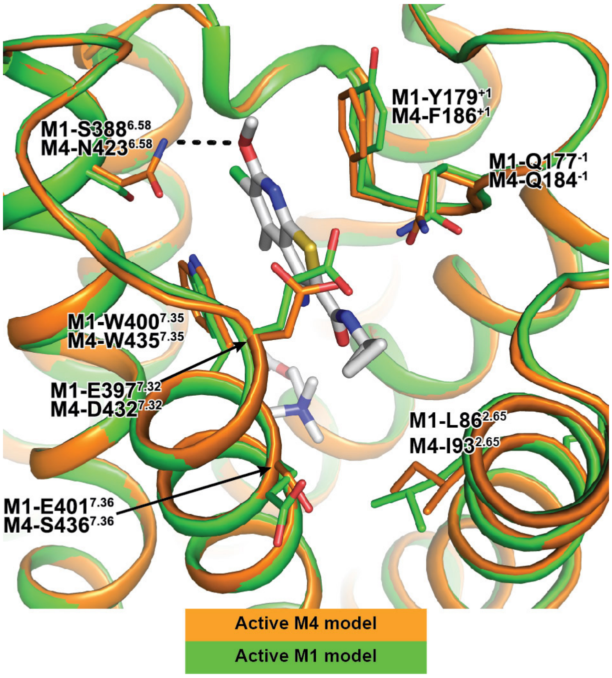
Extended Data Figure 6 | Ligand interaction diagrams for the M4 receptor. a, **b**, The molecular interactions between the (**a**) orthosteric and (**b**) allosteric binding sites are shown by the program MOE[65] for the inactive (M4•tiotropium structure) and active states (M4•acetylcholine•LY2033298 model). Residues with a bold outline were selected in this study or others[20,21] as single-point mutations.



Extended Data Figure 7 | Identification of key residues that govern LY2033298 affinity and binding cooperativity with ACh at the M4 receptor. Competition between a fixed concentration of [^{3}H]QNB and increasing concentrations of ACh (black circles), LY2033298 (blue triangles), or LY2033298 in the presence of an IC_{20} concentration of ACh (red squares) are shown. The curves drawn through the points represent the best global fit of an extended ternary complex model. For data sets where the binding of [^{3}H]QNB changed by less than 10% at 10^{-5} M LY2033298 relative to zero LY2033298, the value of α' was fixed to 1 (connecting line shown). Data points represent the mean ± s.e.m. of at least three experiments performed in triplicate.



Extended Data Figure 8 | Comparison of cooperativity network residues between the inactive and active-states. a, **b**, Chemical structures of (**a**) the M4 ligands used in this study and (**b**) the M2 ligands from the active-state crystal structures (PDB accession number 4MQT and 4MQS). **c**–**f**, Mapping of the allosteric network onto the (**c**, **d**) inactive M4 structure (blue residues), M4 active-state model (orange residues) and (**e**, **f**) the inactive (yellow residues) and active-state M2 structures (magenta and green residues) with views from the (**c**, **e**) membrane or (**d**, **f**) extracellular surface. Ligands are coloured according to element: carbon, cyan; oxygen, red; nitrogen, blue; sulfur, yellow; chlorine, green.



Extended Data Figure 9 | LY2033298 binding to active-state M1 and M4 models. Comparison of active-state M1 (green) and M4 (orange) models bound to LY2033298 and acetylcholine, with acetylcholine and LY2033298 shown as sticks and coloured according to element: carbon, white; oxygen, red; nitrogen, blue; sulfur, yellow; chlorine, green. Several residues surrounding LY2033298 are shown as sticks and coloured according to receptor. M4-N423$^{6.58}$ is predicted to undergo significant movement between the inactive and active states to form a hydrogen bond with the methoxy group of LY2033298. In the M1 receptor this residue is a serine (S388$^{6.58}$) and is unable to form a similar hydrogen bond. However, mutation of N423$^{6.58}$ to alanine at the M4 receptor results in no loss of LY2033298 affinity, but does result in a sixfold loss in cooperativity between acetylcholine and LY2033298 (Supplementary Table 3). This is suggestive of selectivity being derived through cooperativity as a possible mechanism between the M1 and M4 receptors. Additional determinants for M1 and M4 selectivity could also arise through differences in residues on TMs 2 and 7, which contribute to (I93$^{2.65}$) or sit proximal to (D432$^{7.32}$ and S436$^{7.36}$) the allosteric network.

Extended Data Table 1 | Data collection and refinement statistics

Data collection*	M1-T4L•tiotropium	M4mT4L•tiotropium
Beamline	GM/CA 23-ID-D	GM/CA 23–ID-D
Number of crystals	8	64
Space group	P 2_1 2_1 2_1	P 1 2_1 1
Cell dimensions		
a, b, c (Å)	58.0, 72.2, 175.7	48.5, 172.0, 60.7
α, β, γ (°)	90, 90, 90	90, 94.4, 90
Resolution (Å)	30.00–2.70 (2.80–2.70)	28.99–2.60 (2.69–2.60)
R_{merge} (%)	15.5 (83.6)	21.3 (94.7)
<I/σI>	8.7 (1.9)	8.2 (1.5)
$CC_{1/2}$ (%)	N.D.†	99.0 (54.5)
Completeness	97.1 (98.2)	99.7 (97.9)
Multiplicity	4.3 (4.2)	13.1 (4.9)
Refinement		
Resolution (Å)	30.00–2.70	28.53–2.60
No. of reflections working / test set	19,223 / 1,011	30,299 / 2,146
R_{work}/R_{free} (%)	22.8 / 27.5	22.7 / 24.0
No. of atoms (Chain A / Chain B)		
Protein	3439	3058 / 3115
Ligands	210	176 / 167
Average B-factors (Chain A / Chain B; Å^2)	73.6	74.0 / 70.5
Receptor	62.5	63.9 / 64.6
T4 lysozyme	64.4	99.5 / 85.3
Tiotropium	66.5	47.4 / 47.9
Allosteric site PEG 300	–	74.9 / 81.6
Waters	58.1	53.7 / 46.7
Other ligands	74.8	80.8 / 78.4
RMS deviation from ideality		
Bond length (Å)	0.009	0.002
Bond angles (°)	1.39	0.61
Ramachandran statistics‡		
Favored regions (%)	96.2	99.2
Allowed regions (%)	3.6	0.8
Outliers (%)	0.2	0.0

*Highest shell statistics in parenthesis.
†N.D., Not determined, because the structure was solved before $CC_{1/2}$ values were introduced[66].
‡As calculated by Molprobity.

doi:10.1038/nature16978

Positron annihilation signatures associated with the outburst of the microquasar V404 Cygni

Thomas Siegert[1], Roland Diehl[1], Jochen Greiner[1], Martin G. H. Krause[1,2], Andrei M. Beloborodov[3], Marion Cadolle Bel[4], Fabrizia Guglielmetti[1], Jerome Rodriguez[5], Andrew W. Strong[1] & Xiaoling Zhang[1]

Microquasars[1–4] are stellar-mass black holes accreting matter from a companion star[5] and ejecting plasma jets at almost the speed of light. They are analogues of quasars that contain supermassive black holes of 10^6 to 10^{10} solar masses. Accretion in microquasars varies on much shorter timescales than in quasars and occasionally produces exceptionally bright X-ray flares[6]. How the flares are produced is unclear, as is the mechanism for launching the relativistic jets and their composition. An emission line near 511 kiloelectronvolts has long been sought in the emission spectrum of microquasars as evidence for the expected electron–positron plasma. Transient high-energy spectral features have been reported in two objects[7,8], but their positron interpretation[9] remains contentious. Here we report observations of γ-ray emission from the microquasar V404 Cygni during a recent period of strong flaring activity[10]. The emission spectrum around 511 kiloelectronvolts shows clear signatures of variable positron annihilation, which implies a high rate of positron production. This supports the earlier conjecture that microquasars may be the main sources of the electron–positron plasma responsible for the bright diffuse emission of annihilation γ-rays in the bulge region of our Galaxy[11]. Additionally, microquasars could be the origin of the observed megaelectronvolt continuum excess in the inner Galaxy.

Flaring activity of V404 Cygni was discovered with the Swift/BAT and MAXI X-ray monitors[10], and observations with INTEGRAL[12] started within two days, lasting from 17 to 31 June 2015. This recent activity period of V404 Cygni—11 days (15–26 June)—was shorter than its previous outbursts[13] in 1934 and 1989. The source exhibited multiple flares within hours, outshining[13] the brightest persistent X-ray source in the sky, the Crab nebula, by factors of up to 40. V404 Cygni represents a 'gold standard' for multi-wavelength observations, as the parameters of the binary system are well known. It is composed of a 9-solar-mass black hole with a companion star of 0.7 solar masses[14] in a 6.5-day orbit[15] and a 67° inclination angle[14] to the line of sight, and is located at a distance of 2.4 kpc (ref. 16).

INTEGRAL/SPI spectrometer data were extracted and calibrated following the standard procedures, including careful accounting for the detector response and background (see Methods section 'SPI data extraction'). For spectral analysis, the data have been summed into three epochs of ~3 days during the V404 Cygni flaring period. In each of these three epochs (INTEGRAL orbits 1554, 1555 and 1557) we have detected a significant excess of emission around 511 keV, consistent with positron annihilation.

At energies below 200 keV, the spectrum is well described by the standard model[17] of thermal Comptonization plus reflection, but at higher energies a large excess (~18 s.d.) appears above its high-energy tail (see Fig. 1). To further quantify this excess, we added a model spectrum of electron–positron plasma with temperature T (see Methods section 'Spectral fitting', Extended Data Fig. 1 and Extended Data Table 1); the temperature serves as a measure of the annihilation line width.

The characteristic curved spectral shape of pair annihilation emission describes the excess in all three epochs well. Its inclusion leads to a 5-s.d. overall improvement of the fit compared to a model describing the excess with a power law (see Extended Data Tables 2–4 for details, including model alternatives). No shift with respect to the

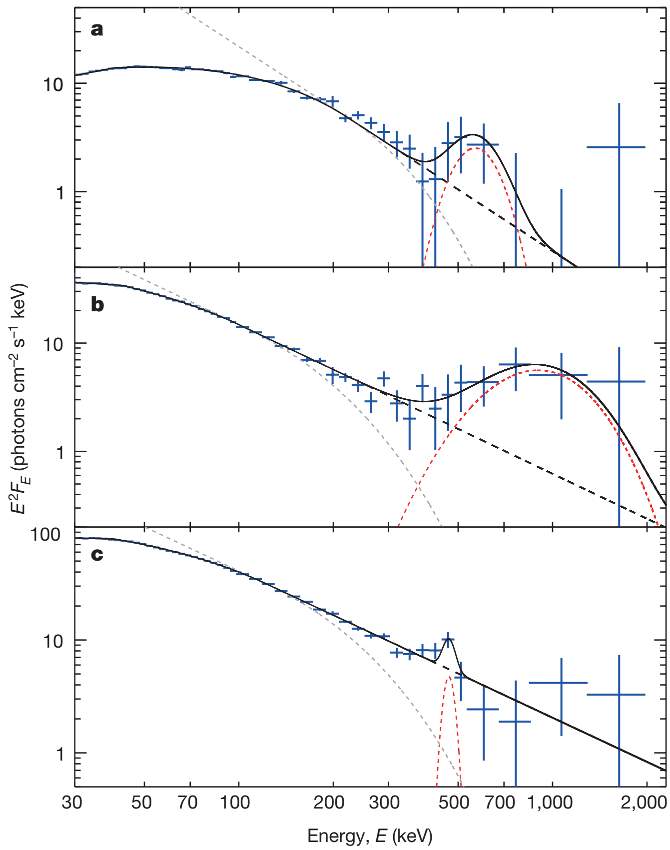


Figure 1 | Spectral evolution of V404 Cygni. a–c, Spectra in the soft γ-ray band in three different flaring epochs (**a–c** show the spectra measured in INTEGRAL orbits 1554, 1555 and 1557, corresponding to epochs 1, 2 and 3, respectively). Data (blue; error bars, 1 s.d.) are fitted as the sum of the Comptonization continuum (black dashed curve) and annihilation radiation from a relativistic hot plasma (red dashed curve). The standard thermal Comptonization model (grey dashed curve) fits the data up to ~200 keV; it declines exponentially at higher energies and falls short of the observed flux. Our conservative, modified continuum model follows a power law instead. E, energy; F_E, flux at energy E.

[1]Max-Planck-Institut für extraterrestrische Physik, Gießenbachstraße 1, 85748 Garching, Germany. [2]Universitäts-Sternwarte München, Ludwig-Maximilians-Universität, Scheinerstraße 1, 81679 München, Germany. [3]Physics Department and Columbia Astrophysics Laboratory, Columbia University, 550 West 120th Street, New York, New York 10027, USA. [4]Max Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany. [5]Laboratoire Astrophysique Instrumentation Modélisation, UMR 7158, CEA/CNRS/Université Paris Diderot, CEA DSM/IRFU/SAp, 91191 Gif-sur-Yvette, France.

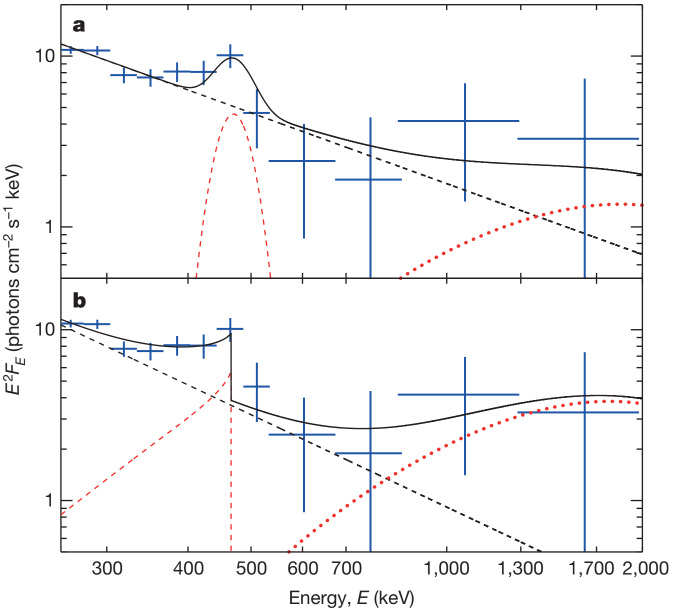


Figure 2 | Model alternatives for epoch 3. Shown is the high-energy part of the spectrum in epoch 3 fitted with different annihilation models. Data are shown blue in both panels (error bars, 1 s.d.). In both panels, the black dashed curve shows the Comptonization continuum, and the black solid line, the total model spectrum. **a**, Two components of annihilation in flight in a thermal plasma: $kT \approx 4$ keV redshifted by ∼10% (red dashed curve), and $kT \approx 500$ keV (red dotted curve). **b**, Positronium three-photon annihilation redshifted by ∼10% (red dashed curve) and annihilation in flight in a hot plasma (red dotted curve).

laboratory frame of the annihilation line is required to fit the data in epochs 1 and 2. A redshift of ∼10% is required in epoch 3. The line width varies strongly between the three epochs—the temperature parameter varies from a few keV to about 200 keV.

Electron–positron pair production is expected near luminous accreting black holes when their radiation spectra extend above $E = m_ec^2 = 511$ keV (refs 18, 19), where m_e is the mass of the electron, and c is the speed of light. Pairs are produced in collisions between MeV γ-rays, with an average cross-section $\sigma_{\gamma\gamma} \approx 10^{-25}$ cm^2. This process is efficient because of the small size of the source, radius $r \approx (3–10)r_g$, where $r_g \approx 10$ km is the gravitational radius of the black hole. During the V404 Cygni flares, the observed luminosity in photons of energy $E \approx m_ec^2$ increases up to $L_1 \approx 10^{37}$ erg s^{-1}, which corresponds to a photon density $n_1 \approx L_1/(\pi r^2 m_e c^3)$, and an optical depth for collisions between soft γ-rays of $\tau_{\gamma\gamma} \approx n_1\sigma_{\gamma\gamma}r$. Gamma-rays of higher energies are absorbed more efficiently as they can interact with the more numerous X-rays of lower energies. The large optical depth $\tau_{\gamma\gamma} \approx 1$ tends to suppress the spectrum of the central compact source at photon energies $E \gg m_ec^2$, consistent with the upper limit on the GeV emission (obtained from analysing public Fermi/LAT data, $F_{GeV} < 10^{-6}$ photons cm^{-2} s^{-1}; see Methods section 'FERMI/LAT data analysis'). Thus, a significant fraction of the luminosity is converted to pair plasma: the plasma is continually created and annihilated inside the source, forming a broad annihilation line of width equivalent to $kT \approx 100$ keV (ref. 19). Photon collisions outside the source create a pair outflow[20]. Pairs annihilating outside the source are in Compton equilibrium with radiation at a temperature of a few tens of keV. Pair outflows have also been invoked to explain the shape of the X-ray spectra of accreting black holes, in particular their flat slopes and the reduced reflection component[21,22].

These expectations are confirmed by our observations of a broad annihilation feature in epochs 1 and 2. The observed flux corresponds to a positron creation rate of $\dot{N}_\pm \approx 10^{42}$ s^{-1}, and to an energy generation rate in the form of pairs, $L_\pm$, of $L_\pm \approx \dot{N}_\pm m_e c^2 \approx 10^{36}$ erg s^{-1}, a few per cent of the source luminosity[13], in agreement with models[20]. Pair plasma annihilating outside the source is then consistent with observations in epoch 1 ($kT \approx 30$ keV), while annihilation inside the source corresponds to the broad line observed in epoch 2 ($kT \approx 170$ keV).

In epoch 3, there is a hint of a steep decline on the blue side of the line that is characteristic of three-photon annihilation of positronium atoms (see Methods section 'Spectral fitting'). Figure 2 shows the data fitted by a redshifted, narrow annihilation feature (Fig. 2a), or alternatively by a positronium-annihilation spectral shape (Fig. 2b). In addition, the broad component extending above 511 keV may be fitted by a very hot annihilation line; this more complex model would correspond to two annihilation regions, cold and hot (see Extended Data Table 5).

The smaller width of the line in epoch 3 ($kT \approx 2–3$ keV), in combination with the detected redshift, poses a challenge to models. The 10% redshift is consistent with a gravitational redshift of the source at radius $r \approx 10r_g$. A line emitted in this region is expected to be significantly broader, owing to the mildly relativistic bulk motions of the plasma and dispersion in gravitational redshift.

The positronium interpretation also poses problems, because it requires[23] a dense plasma with a low temperature, $T < 10^6$ K. The minimum temperature of a source with luminosity $L \approx 10^{38}$ erg s^{-1} and size $r \approx 10r_g \approx 100$ km is its effective blackbody temperature, $T_{eff} \approx [L/(4\pi r^2\sigma_{bb})]^{1/4} \approx 10^7$ K, where σ_{bb} is the Stefan–Boltzmann constant. The accretion flow at larger distances may contain much cooler gas, especially if the central radiation is beamed and shielded by the inner accretion disk. Then the pair plasma created in the central source may be blown out by radiation along the magnetic field lines towards the colder gas, decelerate there and form positronium before annihilating. This scenario is, however, inconsistent with the redshift of the line, in particular when interpreted as a gravitational redshift.

The morphology of the material near the black hole may also be inferred from soft X-ray observations at high spectral resolution[24], suggesting high ionization out to $r \approx 2 \times 10^6$ km. This is a plausible value for the outer edge of the accretion disk. Variability of X-ray line strengths and a changing ionization strongly suggest that most of the X-rays are not received directly but are observed after reflection/reprocessing in the outer region. This is consistent with the flat shape of the hard X-ray continuum[25] up to 100 keV. In this picture, the outer edge of the disk is elevated, blocking our direct view towards the inner disk, and the X-ray continuum variability could be due to changes in the outer disk, rather than to rapid changes in the accretion rate onto the black hole.

The flux evolution of the continuum and annihilation features throughout our observation period is shown in Fig. 3. The two components appear to be correlated. When the X-ray flaring of V404 Cygni faded, the annihilation signal also vanished. However, some annihilation radiation appeared between the X-ray flares. Our time resolution is much longer than the dynamical timescale of the inner accretion disk (milliseconds), the jet ejection timescale[2] (minutes), the light travel time between the two binary components (80 s), and the timescale of adiabatic expansion of ejected blobs[3] (minutes, less than 1 h). Thus, our measurements probe time-averaged values and do not resolve the rise time of the annihilation flux, which may be much shorter than our 6-h bins (see Methods section 'Timing analysis'). Our search for a lag between the X-ray flares and the annihilation radiation was inconclusive, although there is an indication of a lag of several hours (see Methods section 'Timing analysis').

In contrast to hadronic gas, pair plasma is easily accelerated and immediately attains an equilibrium bulk speed $v \approx c/2$ away from the source (governed by the local radiation field anisotropy), forming the base of a relativistic outflow from the accretion disk[20]. The 67° inclination implies a Doppler blueshift of ∼10%, comparable to the gravitational redshift at $r \approx 10r_g$. From our data in epochs 1 and 2, we can neither confirm nor exclude possible residual shifts of the broad line. Annihilation radiation around the source can be strongly affected by the magnetic field configuration, which can change the outflow speed and direction. The power deposited in the pair plasma alone of $L_\pm \approx 10^{36}$ erg s^{-1} would be sufficient to explain the observed typical

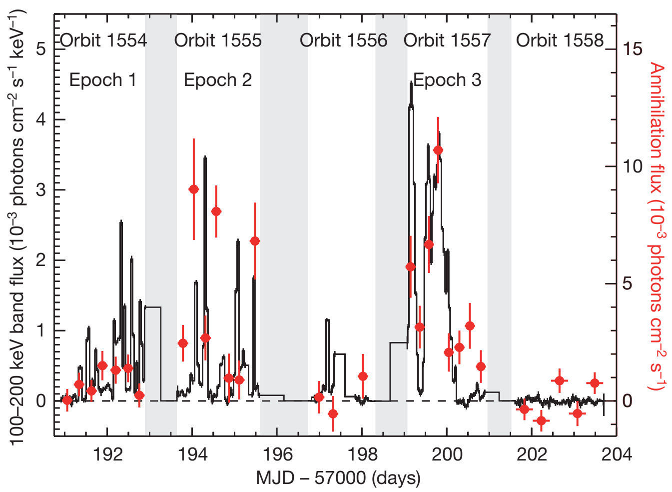


Figure 3 | Time history of X-ray continuum and annihilation emission. The two components of high-energy emission from V404 Cygni are shown as they evolve during the June 2015 flaring period, measured between 17 June and 30 June. INTEGRAL orbits 1554–1558 and epochs 1–3 are shown at the top. The black histogram shows the spectral photon flux in the 100–200-keV band (left-hand *y* axis) and represents the brightness of the Comptonization component emission. The red circles show the photon flux in the annihilation line averaged over 6 h (right-hand *y* axis); vertical error bars, 1 s.d. MJD, modified Julian date. Grey shaded areas mark the regions where no data have been taken.

radio luminosity associated with escaping blobs on larger scales[26]. A changing magnetic field may be responsible for the evolution of the line shape between epochs 1 and 2.

The outflow of pairs from the central source makes microquasars efficient factories, enriching the surrounding medium with positrons. This can help to solve the puzzle of annihilation radiation observed from the bulge region of our Galaxy[11]. In a steady state, the annihilation rate in the inner Galaxy is $(1–2) \times 10^{43}$ positrons s^{-1} and must be balanced by a positron supply with the same rate. During the 10-day flaring period, V404 Cygni produced $\sim 10^{42}$ positrons s^{-1}. In more luminous microquasars such as GRS 1915+105, this number is expected to be even larger. Thus, the steady-state annihilation rate in the inner Galaxy requires only about ten such sources to be active at any time. With duty cycles (that is, flaring versus quiescent epochs) of the order of 10^{-3}, as observed for V404 Cygni, $\sim 10^3–10^4$ accreting black hole binaries of this type would be required in the inner Galaxy, consistent with population synthesis estimates[27].

The excess emission at a few hundred keV to a few MeV that we observed in the flaring epochs of V404 Cygni may also be responsible for the excess (about 10 keV photons $cm^{-2} s^{-1}$ at 1 MeV; ref. 28) found in the diffuse γ-ray continuum spectrum of our Galaxy's ridge[29]. This excess is not accounted for by models of cosmic-ray interactions with interstellar gas[28], nor by inverse-Compton emission, and had been attributed to an unknown source population. A few active sources at an average observed microquasar distance of 5 kpc (ref. 26) would be needed to explain the diffuse MeV excess. The implied duty cycle is consistent with the above estimate for $\sim 10^4$ sources. Thus, the same population of microquasars could also be the origin of the observed inner Galaxy's excess in MeV continuum.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 August 2015; accepted 4 January 2016.
Published online 29 February 2016.

1. Mirabel, I. F., Rodriguez, L. F., Cordier, B., Paul, J. & Lebrun, F. A double-sided radio jet from the compact Galactic Centre annihilator 1E1740.7–2942. *Nature* **358,** 215–217 (1992).
2. Mirabel, I. F. & Rodriguez, L. F. Sources of relativistic jets in the Galaxy. *Annu. Rev. Astron. Astrophys.* **37,** 409–443 (1999).
3. Fender, R. P., Belloni, T. M. & Gallo, E. Towards a unified model for black hole X-ray binary jets. *Mon. Not. R. Astron. Soc.* **355,** 1105–1118 (2004).
4. Mirabel, I. F. & Rodriguez, L. F. A superluminal source in the Galaxy. *Nature* **371,** 46–48 (1994).
5. Remillard, R. A. & McClintock, J. E. X-ray properties of black-hole binaries. *Annu. Rev. Astron. Astrophys.* **44,** 49–92 (2006).
6. Greiner, J., Morgan, E. H. & Remillard, R. A. Rossi X-Ray Timing Explorer observations of GRS 1915+105. *Astrophys. J.* **473,** L107–L110 (1996).
7. Bouchet, L. *et al.* Sigma discovery of variable e^+-e^- annihilation radiation from the near Galactic center variable compact source 1E 1740.7−2942. *Astrophys. J.* **383,** L45–L48 (1991).
8. Goldwurm, A. *et al.* Sigma/GRANAT soft gamma-ray observations of the X-ray nova in Musca — discovery of positron annihilation emission line. *Astrophys. J.* **389,** L79–L82 (1992).
9. Sunyaev, R. *et al.* X-ray nova in Musca (GRS 1124–68): hard X-ray source with narrow annihilation line. *Astrophys. J.* **389,** L75–L78 (1992).
10. Kuulkers, E. *et al.* Renewed activity of V404 Cyg (GS 2023+338). *Astron. Telegr.* **7647** (2015).
11. Prantzos, N. *et al.* The 511 keV emission from positron annihilation in the Galaxy. *Rev. Mod. Phys.* **83,** 1001–1056 (2011).
12. Winkler, C. *et al.* The INTEGRAL mission. *Astron. Astrophys.* **411,** L1–L6 (2003).
13. Rodriguez, J. *et al.* Correlated optical, X-ray, and γ-ray flaring activity seen with INTEGRAL during the 2015 outburst of V404 Cygni. *Astron. Astrophys.* **581,** L9–L13 (2015).
14. Khargharia, J., Froning, C. S. & Robinson, E. L. Near-infrared spectroscopy of low-mass X-ray binaries: accretion disk contamination and compact object mass determination in V404 Cyg and Cen X-4. *Astrophys. J.* **716,** 1105–1117 (2010).
15. Miller-Jones, J. C. A. *et al.* The first accurate parallax distance to a black hole. *Astrophys. J.* **706,** L230–L234 (2009).
16. Casares, J., Charles, P. A. & Naylor, T. A 6.5-day periodicity in the recurrent nova V404 Cygni implying the presence of a black hole. *Nature* **355,** 614–617 (1992).
17. Done, C., Gierlinski, M. & Kubota, A. Modelling the behaviour of accretion flows in X-ray binaries. *Astron. Astrophys. Rev.* **15,** 1–66 (2007).
18. Maciolek-Niedzwiecki, A., Zdziarski, A. A. & Coppi, P. S. Electron/positron pair production and annihilation spectral features from compact sources. *Mon. Not. R. Astron. Soc.* **276,** 273–282 (1995).
19. Svensson, R. Non-thermal pair production in compact X-ray sources — first-order Compton cascades in soft radiation fields. *Mon. Not. R. Astron. Soc.* **227,** 403–451 (1987).
20. Beloborodov, A. M. Electron-positron outflows from gamma-ray emitting accretion discs. *Mon. Not. R. Astron. Soc.* **305,** 181–189 (1999).
21. Fabian, A. C. *et al.* Properties of AGN coronae in the NuSTAR era. *Mon. Not. R. Astron. Soc.* **451,** 4375–4383 (2015).
22. Beloborodov, A. M. Plasma ejection from magnetic flares and the X-ray spectrum of Cygnus X-1. *Astrophys. J.* **510,** L123–L126 (1999).
23. Crannell, C. J., Joyce, G., Ramaty, R. & Werntz, C. Formation of the 0.511 MeV line in solar flares. *Astrophys. J.* **210,** 582–592 (1976).
24. King, A. L., Miller, J. M., Raymond, J., Reynolds, M. T. & Morningstar, W. High-resolution *Chandra* HETG spectroscopy of V404 Cygni in outburst. *Astrophys. J.* **813,** L37 (2015).
25. Natalucci, L. *et al.* High energy specral evolution of V404 Cygni during the 2015 June outburst as observed by *INTEGRAL*. *Astrophys. J.* **813,** L21 (2015).
26. Gallo, E., Fender, R. P. & Pooley, G. G. A universal radio-X-ray correlation in low/hard state black hole binaries. *Mon. Not. R. Astron. Soc.* **344,** 60–72 (2003).
27. Sadowski, A., Ziółkowski, J., Belczyński, K. & Bulik, T. Calculations of the Galactic population of black hole X-ray binaries. *AIP Conf. Proc.* **1010,** 404–406 (2008).
28. Grenier, I. A., Black, J. H. & Strong, A. W. The nine lives of cosmic rays in galaxies. *Annu. Rev. Astron. Astrophys.* **53,** 199–246 (2015).
29. Strong, A. W. *et al.* Gamma-ray continuum emission from the inner Galactic region as observed with INTEGRAL/SPI. *Astron. Astrophys.* **444,** 495–503 (2005).

Acknowledgements The INTEGRAL/SPI project has been completed under the responsibility and leadership of CNES; we are grateful to ASI, CEA, CNES, DLR, ESA, INTA, NASA and OSTC for support of this ESA space science mission. R.D. and J.G. are also supported by the Munich excellence cluster 'Origin and evolution of the Universe'. M.G.H.K. is supported by the Deutsche Forschungsgemeinschaft, project number PR 569/10-1, as part of DFG Priority Program 1573. J.R. acknowledges funding support from the French Research National Agency, CHAOS project ANR-12-BS05-0009, and from the UnivEarthS Labex program of Sorbonne Paris Cité.

Author Contributions T.S. was responsible for the spectroscopy analysis, data modelling, and paper writing, and R.D. led the analysis and paper writing. J.G., M.G.H.K. and A.M.B. were responsible for interpretational aspects and crucial inputs to the paper. J.G. was responsible for analysis of the Fermi data, M.C.B and J.R. contributed in microquasar physics, F.G. in data analysis, A.W.S. in γ-ray continuum and cosmic-ray physics, and X.Z. was responsible for data preparation and reduction, and the instrument response.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.S. (tsiegert@mpe.mpg.de).

METHODS

SPI data extraction. INTEGRAL's[12] SPI γ-ray spectrometer[30,31] provides photon event measurements, where pulse heights are recorded in a 19-element germanium semiconductor detector array. A coded mask imprints a shadowgram of celestial sources onto the detector array, as particular regions of the sky are blocked from detectors through the opaque tungsten elements of the mask. Re-orientations of the telescope by small angles ('dithering', $\sim$2°) provide an additional coding pattern. Coded-mask imaging allows sources within the telescope field of view of $\sim$30° to be resolved with a precision of $\sim$2°.

We accumulate event data for each telescope pointing and detector and bin them into spectra, after initial preprocessing with INTEGRAL standard OSA 9 software. We use single-detector hits only, to ensure a well-known spectral detector response. Above 530 keV, we use pulse-shape filtered events to suppress an electronics malfunction that contaminates normal event data[32], and we apply their appropriate filter efficiency factor of 0.8. Pointings typically lasted one hour during the observations of V404 Cygni.

For adequate spectral precision, we accumulate three-day orbit data to extract a spectrum for V404 Cygni. The full set of detector spectra over an orbit is then fitted with parameterized models for instrumental background and for all candidate sources in the field of view[33]. The celestial emission models are folded into the data space of spectra using the coded mask shadowing properties, and the energy response of the detectors. Thus, intensity scaling factors per energy bin are obtained for all candidate sources within the telescope field of view, which are V404 Cyg, Cyg X-1, Cyg X-3, Cyg A, 3A1954+319 and EXO2030+375. Our model fit of background and sky contributions to data then provides spectra for all these sources.

The background is derived from data of the two preceding INTEGRAL orbits (1552, 1553). Being from the pre-flaring periods of V404 Cygni and being pointed to regions devoid of sources at high Galactic latitudes, these data are free of any high-energy (>300 keV) signals from the sky. The background is modelled as a constant detector-intensity ratio (pattern) per energy bin, taken from these independent data. Owing to the strong time variability, the overall amplitudes per energy bin (background-intensity scaling factors) have been determined on a one-hour timescale.

The continuum sensitivity (3 s.d.)[31] around 400 keV for an exposure of two days (corresponding to the dead-time corrected on-time of one orbit) is about 3.0×10^{-6} photons $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mathrm{keV}^{-1}$ with only a weak dependence towards higher energies (about 1.8×10^{-6} photons $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mathrm{keV}^{-1}$ at 1,600 keV). This is sufficient to detect sources with intensities of $\sim$100 mCrab above 400 keV. V404 Cygni, for example, shows an average intensity of $(12.0 \pm 3.5) \times 10^{-6}$ photons $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mathrm{keV}^{-1}$ around 500 keV, and about $(2.5 \pm 1.6) \times 10^{-6}$ photons $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mathrm{keV}^{-1}$ around 1,000 keV, during its flaring epochs (see Fig. 1), in accordance with sensitivity limits. Thus, at higher energies, statistical uncertainties dominate over systematics.

Some cross-talk among the sources is expected to occur in the data of a coded-mask instrument with a large field-of-view. Also, concerns about artefacts from background model inadequacies need special attention. We therefore compare the spectra of a set of hypothetical sources fitted additionally in the empty regions of the field-of-view, together with the above-mentioned known sources, as commonalities or anti-correlations among them would be a hint of background or cross-talk issues. For a given data set, the total number of measured photons is fixed, so that the inclusion of test sources may result in an increased flux value for one source while at the same time reduce the flux for another (anti-correlations). But apart from clear signals in the spectra of the real sources, all source spectra in the empty field regions are fully consistent with background only, and in particular do not show any spectral signatures beyond statistical fluctuations.

Spectral fitting. The spectra for the three flaring epochs were fitted as a sum of two components: a Comptonized continuum $C(E)$ and pair annihilation emission $P(E)$

$$f(E) = C(E) + P(E) \tag{1}$$

Continuum X-ray spectra of accreting black holes in the hard state are well explained by the thermal Comptonization model[17,34]: seed low-energy photons (for example, thermal radiation from the optically thick accretion disk) are re-processed by a hot plasma corona through Compton scattering to higher energies, leading to a power-law spectrum. The power law cuts off exponentially where the photon energy exceeds the mean energy of the scattering electrons, which is typically $\sim$50–100 keV (refs 17, 34). This model describes the spectra of V404 Cygni reasonably well below $\sim$200 keV, between and during flaring[13]. However, additional emission is required above 200 keV. One can model this emission as an additional power-law component due to scattering by non-thermal, accelerated particles. We use the following approximation to the Comptonization spectrum

$$C(E) = \begin{cases} A_0\left(\dfrac{E}{E_0}\right)^{\alpha}, & E < E_C \\ A_0\left(\dfrac{E}{E_0}\right)^{\alpha}\exp\left(\dfrac{E_C - E}{E_F}\right), & E_C \le E \le E_X \\ B_0\left(\dfrac{E}{E_0}\right)^{\beta}, & E > E_X \end{cases} \tag{2}$$

with $B_0 = A_0\left(\frac{E_X}{E_0}\right)^{\alpha-\beta}\exp\left(\frac{E_C - E_X}{E_F}\right)$, where $E_0 = 100$ keV is a normalization convention.

The model of thermal pair annihilation (TPA)[19,35], $P(E)$, is calculated as follows. Electron and positron energy distributions are described by a Maxwell–Jüttner distribution, which is the relativistic Maxwell–Boltzmann distribution

$$f(\gamma) = \frac{\gamma^2\beta}{\theta K_2(1/\theta)}\exp\left(-\frac{\gamma}{\theta}\right) \tag{3}$$

Here $\gamma = (1-\beta^2)^{-1/2}$ is the Lorentz factor, $\beta = v/c$ is the dimensionless velocity, $\theta = \frac{kT}{mc^2}$ is the dimensionless temperature of the pair plasma with Boltzmann's constant k, electron (positron) mass m, the speed of light c, and K_2 is the modified Bessel function of second kind. The spectral distribution of annihilation photons can be calculated in terms of a dimensionless photon energy $x = \frac{h\nu}{mc^2}$ as[35]

$$\begin{aligned} \frac{\mathrm{d}n}{\mathrm{d}t}(x,\theta)\mathrm{d}x = n_+ n_- c\mathrm{d}x \frac{2}{\theta K_2(1/\theta)^2}\exp\left(-\frac{x}{\theta}\right) \\ \int_1^{\infty} \mathrm{d}s 2(s-1)\sigma_{\mathrm{ann}}(s)\exp\left(-\frac{s}{x\theta}\right) \end{aligned} \tag{4}$$

Here, $n_{\pm}$ are the number densities of the positrons/electrons, $s = x_{\mathrm{cm}}^2$ is the photon momentum in the centre-of-momentum frame, and $\sigma_{\mathrm{ann}}(s)$ is the cross-section for the annihilation process. We have used simplified expressions for the temperature and energy regions of interest[35] which are accurate to 0.04% across the range presented in this analysis.

The TPA model ($P(E)$, equation (4)), shown in Extended Data Fig. 1, does not include possible Doppler shifts due to the bulk motion of the pair plasma or the gravitational redshift, which may be significant when the emission originates near the black hole.

The fitted parameters in our composite model, $f(E)$, are the normalization of continuum at 100 keV, that is, the amplitude A_0, the low-energy power-law index α, the cutoff energy E_C, the e-folding energy E_F, the high-energy power-law index β, the extrapolation energy E_X, the amplitude of the annihilation feature n_+n_-c, and the temperature of the pair plasma T_{TPA}. The parameters E_0 and B_0 do not influence the fit and are only introduced for convenience and clear arrangement.

From these parameters, we calculate the following derived parameters. I_{TPA} is the differential flux integrated over the energy

$$I_{\mathrm{TPA}} = \int_{-\infty}^{+\infty} P(E)\mathrm{d}E \tag{5}$$

The plasma temperature T_{TPA} is multiplied by the Boltzmann constant, k, for a conversion into keV units. Fitted and derived parameters for all epochs can be found in Extended Data Tables 1, 3 and 5.

The fit quality obtained for our one-orbit spectra is not satisfactory at first glance, with χ^2 values of 63.7, 69.2 and 280.9 for 38 (37) d.o.f. (degrees of freedom) (see Extended Data Table 2). Typically, an observation is dominated by a large instrumental background. But below $\sim$200 keV, the number of photons detected from V404 Cygni is very high, (about five times stronger than the background), thus systematics of the instrument response limit the accuracy of the fit. At energies above $\sim$150 keV, our composite model obtains a satisfactory fit quality with χ^2 values of 13.0 to 14.5 for 14 d.o.f. Overall, we detect a large high-energy excess above the conventional Comptonized cut-off description, with a total significance of $\sim$18 s.d., consistent with high-energy excesses reported in other studies[13,25,36]. Our more conservative analysis uses a phenomenological model of the continuum that avoids the high-energy cut-off of the Comptonization model by modifying it to a power-law extension towards high energies. When we use this modified Comptonization description, and add a model spectrum of electron–positron plasma of a temperature T, we find a significance of 5 s.d. altogether (see Extended Data Table 4) for the pair annihilation component above the Comptonization component. Significances have been estimated by χ^2 goodness-of-fit tests, in which extra model components are zeroed in order to evaluate the improvement

in χ^2, compared to the simpler model with fewer parameters. The test statistics therefore follow a χ^2-distribution with 1 d.o.f., in which data are properly scaled by their statistical uncertainties which account for Poissonian fluctuations of source and background.

The fit shown in Fig. 2b assumes a different annihilation model, more typical when positrons are slowed down by, for example, Coulomb interactions before annihilating[37]. This has been well measured in our Galaxy[11], and in terrestrial laboratory experiments[38]. When positrons annihilate at thermal energies, the spectra of positron annihilation show the intermediate formation of a positronium atom, consisting of a positron and an electron. This process is very efficient at low energies and in particular below a threshold energy of 6.8 eV (equivalent to temperatures below 7.8×10^4 K). Positron annihilation then can occur from a singlet state of positronium with two photons at 511 keV (*para*-positronium), or a triplet state with a three photon annihilation continuum spectrum[37] rising in intensity up to a maximum at 511 keV (*ortho*-positronium).

We define the fitted parameters of the additional components as follows. E_{centroid} is the Doppler-shifted peak position of the low-energy annihilation feature, that is, pair plasma annihilation in Fig. 2a, and the sharp edge of the *ortho*-positronium shape, $O(E)$, in Fig. 2b. The plasma temperature kT_{TPA} is only given for cases in which a thermal pair annihilation model has been fitted to the data. In both cases, the additional components are scaled/normalized by a fitted amplitude O_0.

We derive the flux, F_{511}, of the second features by integrating the differential spectral shape over the energy. For Fig. 2a, see equation (5), for the *ortho*-positronium feature, the flux is given by

$$F_{511} = \int_{-\infty}^{+\infty} O(E)\mathrm{d}E = O_0(\pi^2 - 9) \qquad (6)$$

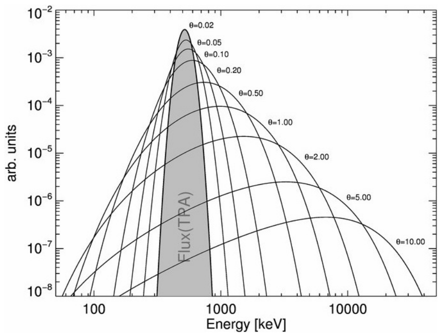
Timing analysis. For the timing analysis of the annihilation emission, we only fit the amplitudes of our model components, A_0 of the continuum, and n_+n_-c of the annihilation feature (see equations (2) and (4) and Fig. 3), in hourly time bins, while holding the spectral shape, that is, all other parameters in both continuum and annihilation model, fixed. This assumes a constant thermal pair annihilation temperature T_{TPA} during one epoch and sums all emission which is spectroscopically identified as due to positron annihilation. The rise time of the excess flux is estimated as a factor of two increase in the positron annihilation component flux. This is difficult to estimate, as it is limited by photon statistics; we estimate a flux doubling within $\sim$2 h. This interprets all fitted signatures on top of the continuum, $C(E)$, as pair plasma emission, rather than accelerated particle emission represented by a power law. In particular, above 500 keV, this may incur a bias towards high annihilation flux values, compared to the detailed spectroscopic analysis per epoch.

The time evolution of the hard X-ray continuum emission (100–200 keV) is obtained[13] from fitting in one-hour intervals the coded-mask response to all sources in the field together with the background model, adjusting the background normalization coefficient per hour.

The linear Pearson correlation coefficient between the 100–200 keV band and the annihilation flux of V404 Cygni during MJD 57191 and 57203 in hourly time bins is 0.45 for zero lag, and 0.34 for a lag of $\sim$15 h.

FERMI/LAT data analysis. We have analysed the all-sky survey data of the FERMI/LAT instrument taken during the 15.5-h time window when we see the largest positron flare, in epoch 3 at orbit 1557, corresponding to MJD 57199.616 to 57200.261. The Pass8 data of a 20° circle around V404 Cygni in the energy range 100 MeV to 10 GeV have been retrieved from the Fermi Science Support Center (http://fermi.gsfc.nasa.gov/cgi-bin/ssc/LAT/LATDataQuery.cgi; accessed 6 August 2015). We employed the unbinned likelihood analysis as implemented in the user-contributed LATAnalysisScripts (http://fermi.gsfc.nasa.gov/ssc/data/analysis/scitools/LATAnalysisScripts.html; accessed 6 August 2015) provided by FSSC, after changing to Pass8 details. Standard event cleaning/removal was applied (http://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_Data_Exploration/Data_preparation.html; accessed 6 August 2015). We used the gll_iem_v06.fits Galactic interstellar emission model, and the iso_P8R2_SOURCE_V6_v06.txt isotropic spectral template for the SOURCE event class (front+back). The Cygnus Loop, Cygnus Cocoon, γ Cyg and HB 21 have been included in the fitting, as well as bright 3FGL sources. No source is detected at the position of V404 Cygni. We derive upper limits of 8×10^{-7} photons cm^{-2} s^{-1} in the 100 MeV to 1 GeV band, and 3×10^{-9} photons cm^{-2} s^{-1} in the 1–10 GeV band.

30. Vedrenne, G. *et al.* SPI: the spectrometer aboard INTEGRAL. *Astron. Astrophys.* **411,** L63–L70 (2003).
31. Roques, J.-P. *et al.* SPI/INTEGRAL in-flight performance. *Astron. Astrophys.* **411,** L91–L100 (2003).
32. Weidenspointner, G. *et al.* First identification and modelling of SPI background lines. *Astron. Astrophys.* **411,** L113–L116 (2003).
33. Bouchet, L. *et al.* INTEGRAL SPI all-sky view in soft gamma rays: a study of point-source and Galactic diffuse emission. *Astrophys. J.* **679,** 1315–1326 (2008).
34. Malzac, J. On the nature of X-ray corona of black hole binaries. *Int. J. Mod. Phys. Conf. Ser.* **08,** 73–83 (2012).
35. Svensson, R., Larrson, S. & Poutanen, J. A simple formula for the thermal pair annihilation line emissivity. *Astron. Astrophys.* **120c** (Suppl.), 587–590 (1996).
36. Roques, J.-P. *et al.* First INTEGRAL observations of V404 Cygni during the 2015 outburst: spectral behavior in the 20–650 keV energy range. *Astrophys. J.* **813,** L22 (2015).
37. Ore, A. & Powell, J. L. Three-photon annihilation of an electron-positron pair. *Phys. Rev.* **75,** 1696–1699 (1949).
38. Sharma, S. C. & McNutt, J. D. Positron annihilation in gaseous nitrogen. *Phys. Rev. A* **18,** 1426–1434 (1978).



Extended Data Figure 1 | Spectral shape of annihilation emission from a relativistic thermal pair plasma. Each curve shows intensity per unit energy and is labelled with θ ($= kT/m_e c^2$), the dimensionless temperature. The model is used to quantify the width of the observed annihilation line (see Methods section 'Spectral fitting'). The grey shaded area is the integrated flux over all energies for this thermal plasma annihilation model (TPA).

Extended Data Table 1 | Spectral fit parameters for the flaring epochs of V404 Cygni

Orbit	A_0 $\left[\frac{10^{-3}\text{ph}}{\text{cm}^2\ \text{s keV}}\right]$	α	E_C [keV]	E_F [keV]	B_0 $\left[\frac{10^{-3}\text{ph}}{\text{cm}^2\ \text{s keV}}\right]$	β	E_X [keV]	I_{TPA} $\left[\frac{10^{-3}\text{ph}}{\text{cm}^2\ \text{s}}\right]$	kT_{TPA} [keV]
1554	2.31(11)	-1.44(5)	40(2)	91(5)	2.18(59)	-3.89(9)	220(113)	1.9(1.2)	29(14)
1555	2.91(16)	-2.18(5)	42(1)	86(3)	1.48(3)	-3.38(4)	95(12)	6.5(1.6)	173(46)
1557	7.30(5)	-2.08(1)	40(1)	99(1)	4.08(3)	-3.30(2)	122(7)	1.2(9)	2(1)

These epochs are shown in Fig. 1. Uncertainties are given in brackets, in units of the last digit. The redshift seen in orbit 1557 is 0.101(1). The amplitudes A_0 and B_0 are normalized to the flux at 100 keV. For nomenclature, see Methods section 'Spectral fitting'.

Extended Data Table 2 | Goodness-of-fit for spectra

Orbit	*1554*	*1555*	*1557*
χ^2(M1) / dof (total spec.)	63.7 / 38	69.2 / 38	280.9 / 37
χ^2(M1) / dof (E < 150 keV)	49.2 / 24	56.2 / 24	265.7 / 24
χ^2(M1) / dof (E > 150 keV)	14.5 / 14	13.0 / 14	15.2 / 13
χ^2(M2) / dof (E > 150 keV)	12.5 / 16	25.6 / 16	30.1 / 16

The goodness-of-fit for spectra corresponding to epochs shown in Fig. 1 is measured by χ^2/d.o.f. (d.o.f., degrees of freedom). We illustrate systematics by giving results for different energy regions, and for different models (see Methods section 'Spectral fitting'). Model 1 (M1) is the full thermal pair annihilation model, and model 2 (M2) is a high-energy cut-off power-law model for Comptonized disk emission plus another independent power law capturing accelerated particle emission.

Extended Data Table 3 | Spectral fit result details for the alternative model M2

Orbit (Model)	A_{01} $\left[\frac{10^{-3}\mathrm{ph}}{\mathrm{cm}^2\ \mathrm{s\ keV}}\right]$	α_1	E_{C1} [keV]	E_{F1} [keV]	A_{02} $\left[\frac{10^{-3}\mathrm{ph}}{\mathrm{cm}^2\ \mathrm{s\ keV}}\right]$	α_2
1554 (M2)	2.26(15)	-1.36(3)	40(2)	79(4)	0.14(6)	-1.94(33)
1555 (M2)	2.58(24)	-2.16(6)	42(1)	70(3)	0.37(3)	-2.27(76)
1557 (M2)	7.38(25)	-1.50(19)	38(1)	50(10)	1.86(43)	-2.62(20)

See also Extended Data Table 2. Model 2 (M2) is as in Extended Data Table 2. The amplitudes A_{01} and A_{02} are normalized to the flux at 100 keV. For nomenclature, see Methods section 'Spectral fitting'.

Extended Data Table 4 | Significance estimates of additional components

Orbit	*1554*	*1555*	*1557*	*Total*
$\Delta\chi_1^2$ / Significance [σ]	5.7 / 2.4	80.3 / 9.0	250.8 / 15.8	336.8 / 18.3
$\Delta\chi_2^2$ / Significance [σ]	3.5 / 1.9	12.4 / 3.5	8.2 / 2.9	24.1 / 4.9

$\Delta\chi_1^2$ describes the improvement of an additional high-energy extension over a high-energy cut-off power-law model only, and $\Delta\chi_2^2$ of an additional thermal pair annihilation feature over the model combining a high-energy cut-off power-law model and its high-energy power-law extension. Significances have been calculated by χ^2-tests with one additional component each. For nomenclature, see Methods section 'Spectral fitting'.

Extended Data Table 5 | Spectral fit parameters for the third flaring epoch

Orbit	B_0 $\left[\frac{10^{-3}\mathrm{ph}}{\mathrm{cm}^2\ \mathrm{s\ keV}}\right]$	β	E_X [keV]	I_{TPA} $\left[\frac{10^{-3}\mathrm{ph}}{\mathrm{cm}^2\ \mathrm{s}}\right]$	kT_{TPA} [keV]	F_{511} $\left[\frac{10^{-3}\mathrm{ph}}{\mathrm{cm}^2\ \mathrm{s}}\right]$	$E_{centroid}$ [keV]	kT_{TPA} [keV]
1557a	4.27(42)	-3.4(1)	<200	1.2(1.0)	511(47)	1.5(0.5)	458(25)	4(3)
1557b	5.73(1.20)	-3.8(3)	<200	3.3(2.9)	489(26)	5.3(1.9)	466(-3/+45)	-

The energy region between 200 keV and 2,000 keV for orbit number 1557 is shown. In addition to a hot thermal annihilation plasma with temperatures around 500 keV, components are either another, Doppler-shifted, thermal annihilation line (1557a), or an *ortho*-positronium continuum (1557b). The amplitude B_0 is normalized to the flux at 100 keV. The χ^2 goodness-of-fit values are 9.3 and 5.9 for 9 and 8 d.o.f., respectively. For nomenclature, see Methods section 'Spectral fitting'.

doi:10.1038/nature17171

Late Tharsis formation and implications for early Mars

Sylvain Bouley[1,2], David Baratoux[3,4], Isamu Matsuyama[5], Francois Forget[6], Antoine Séjourné[1], Martin Turbet[6] & Francois Costard[1]

The Tharsis region is the largest volcanic complex on Mars and in the Solar System. Young lava flows cover its surface (from the Amazonian period, less than 3 billion years ago) but its growth started during the Noachian era (more than 3.7 billion years ago). Its position has induced a reorientation of the planet with respect to its spin axis (true polar wander, TPW), which is responsible for the present equatorial position of the volcanic province. It has been suggested that the Tharsis load on the lithosphere influenced the orientation of the Noachian/Early Hesperian (more than 3.5 billion years ago) valley networks[1] and therefore that most of the topography of Tharsis was completed before fluvial incision. Here we calculate the rotational figure of Mars (that is, its equilibrium shape) and its surface topography before Tharsis formed, when the spin axis of the planet was controlled by the difference in elevation between the northern and southern hemispheres (hemispheric dichotomy). We show that the observed directions of valley networks are also consistent with topographic gradients in this configuration and thus do not require the presence of the Tharsis load. Furthermore, the distribution of the valleys along a small circle tilted with respect to the equator is found to correspond to a southern-hemisphere latitudinal band in the pre-TPW geographical frame. Preferential accumulation of ice or water in a south tropical band is predicted by climate model simulations of early Mars applied to the pre-TPW topography. A late growth of Tharsis, contemporaneous with valley incision, has several implications for the early geological history of Mars, including the existence of glacial environments near the locations of the pre-TPW poles of rotation, and a possible link between volcanic outgassing from Tharsis and the stability of liquid water at the surface of Mars.

The Tharsis bulge is the largest volcano-tectonic centre on Mars. Its growth started during the Noachian epoch (>3.7 billion years ago)[2] and the associated enormous transfer of mass, energy and release of volatiles from the mantle had implications for the planet's evolution, including its climate, surface environment and mantle dynamics. The earliest signs of activity are limited to Noachian extensional tectonics observed around Claritas Fossae in the ancient Syria Planum, Thaumasia and Tempe Terra regions[2]. Assuming a 100-km-thick elastic spherical shell, models of the topographic effect of the Tharsis load have suggested that most of the bulge was largely in place by the end of the Noachian epoch. If so, it could have influenced the orientation of valley networks[1]. However, the elastic lithosphere thickness at the Noachian according to modern estimates[3] (<30 km) corresponds to weak support of the Tharsis load by lithospheric flexure. In fact, early completion of the Tharsis bulge is questionable in light of recent studies: the formation of radial and concentric structures around Tharsis (including Valles Marineris) indicates complex and multi-stage growth of the bulge from the late Noachian to the Amazonian[2]; the inferred crust thickness at the time of formation of the Coracis Fossae rift is evidence of late (<3.5 billion years ago) crustal growth in the Tharsis region[4]; and tectonic structures initially assigned to the Noachian era[2] eventually also affect Hesperian units (Extended Data Fig. 1). Thaumasia and Claritas Fossae are the only Noachian units of Tharsis above the average elevation of the southern hemisphere of Mars (Extended Data Fig. 1). However, these regions at the southern edge of Tharsis have remnant crustal magnetization, in contrast with the rest of the magmatic province, and their formation has been related to a pre-Tharsis orogenic event[5].

The rotation axis orientation of a planet is constrained by its degree-2 gravity field, which is directly related to non-spherical contributions to the inertia tensor. Before Tharsis, the hemispheric dichotomy of the topography controlled the orientation of the planet relative to its spin axis[6]. The growth of Tharsis must have induced a reorientation of the planet with respect to its spin axis (TPW) moving the mass anomaly to its current position close to the equator[6–10]. The rotation poles before Tharsis' formation and TPW (palaeo poles) can be constrained by removing the Tharsis and remnant rotational figure contributions to the degree-2 gravity field. This yields a palaeo north pole at ($100.5° \pm 49.5°$ W, $71.1°^{+17.5°}_{-14.4°}$ N) and a palaeo south pole at ($79.5° \pm 49.5°$ E, $71.1°^{+17.5°}_{-14.4°}$ S)[10], near the centre of the crustal dichotomy (80° E, 60° S)[9].

In addition to a rotation of the geographic reference, TPW induces an adjustment of the topography as a result of a new equilibrium rotational figure[10]. A large TPW event (20°–25°) such as the one triggered by the formation of Tharsis may have had global implications. For instance, the distribution of surface landforms influenced by local slopes or early Mars climate simulations should be calculated within a reference frame corresponding to the planet's configuration at the time of the formation of such landforms. Valley networks, incised during an early climate distinct from the present climate, are mainly located on the highlands[11,12] within a domain of latitudes ranging from −60° S to +30° N (Fig. 1a). In the present reference frame, the density and spread in latitude of Early Hesperian/Noachian valley networks show large variations with longitude[12]. The valley network distribution is dominated by local or regional slopes rather than by elevations[12,13]. However, no combination of topographic parameters is able to explain the lack of valley networks between 30° S and 60° S from the Argyre basin to the east of the Hellas basin and the very low density in Arabia Terra (Fig. 1). In fact, the valley networks occur in a band that appears to follow a small circle tilted with respect to the present equator. We have quantitatively confirmed this observation by least-squares adjustment of the valley network density map to a small circle (Methods). The normal vector of the plane containing the small circle defines a north pole located north of Tharsis along the meridian passing through the centre of the Tharsis bulge (118° W, 69° N) and a south pole south of the Hellas basin (62° E, 69° S). This rotation axis is consistent with the calculated palaeo north pole location before the

[1]GEOPS—Géosciences Paris Sud, Université Paris-Sud, CNRS, Université Paris-Saclay, Rue du Belvédère, Bâtiment 504-509, 91405 Orsay, France. [2]Institut de Mécanique Céleste et de Calcul des Ephémérides, UMR8028, 77 Avenue Denfert Rochereau, 75014 Paris, France. [3]Geosciences Environnement Toulouse, Université de Toulouse III UMR 5563, 14 Avenue Edouard Belin, 31400 Toulouse, France. [4]Institut de Recherche pour le Développement et Institut Fondamental d'Afrique Noire, Dakar, Sénégal. [5]Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA. [6]Laboratoire de Météorologie Dynamique, Institut Pierre Simon Laplace, CNRS, Université Pierre et Marie Curie, Paris, France.

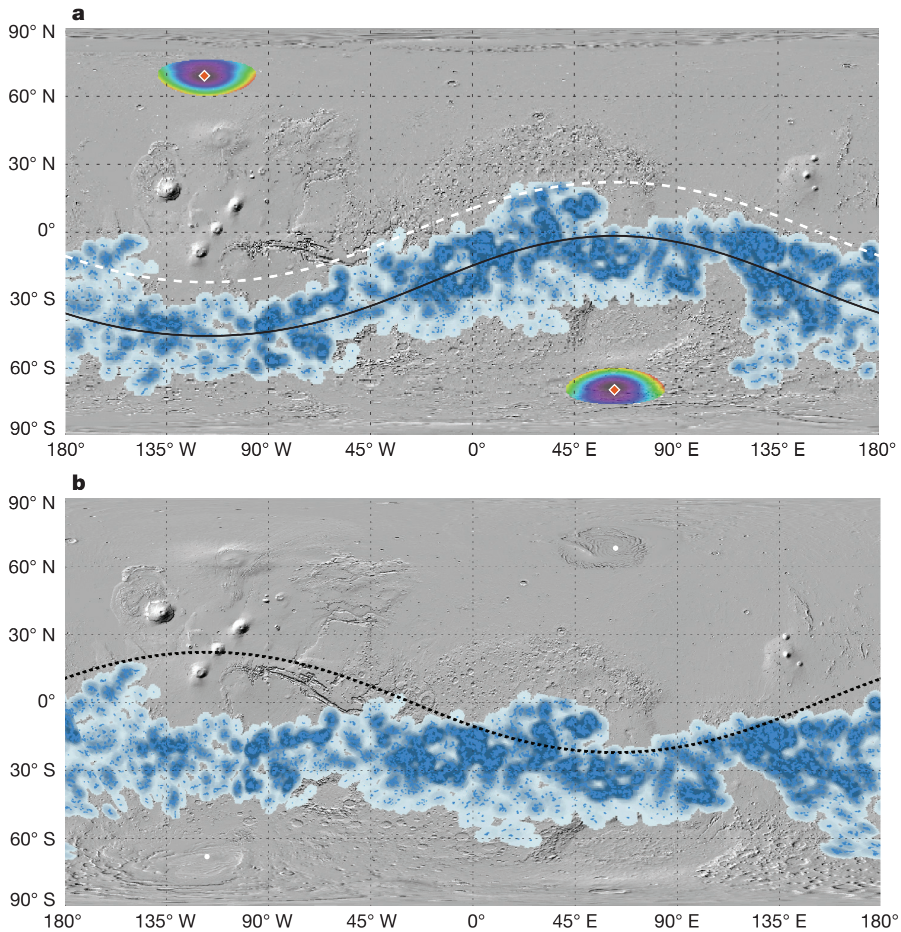


Figure 1 | Noachian/Early Hesperian valley networks distribution and density[12] before and after TPW. a, In the present reference frame. The palaeo pole positions are indicated with a diamond, with the spread in longitude and latitude for solutions shaded from black (14°) to red (15°), corresponding to the associated root-mean-square value for each solution, as given by equation (16) (see Methods). The white dashed line is the pre-TPW equator. **b**, In the pre-TPW configuration. Valley networks occur within a latitudinal band of ±14° centred at 24° S. The black dashed line is the present equator.

formation of Tharsis and TPW ($100.5° \pm 49.5°$ W, $71.1°^{+17.5°}_{-14.4°}$ N; ref. 10). The proximity of the palaeo north pole inferred from the valley network distribution and the one inferred from degree-2 spherical harmonic gravity without Tharsis[10] is remarkable because these results are independent.

In the pre-TPW configuration, the valley network distribution defines a latitudinal band of ±14° centred at 24° S, that is, in the south tropical regions (Fig. 1b). The incision of valley networks spans the Hesperian–Noachian (>3.5 billion years ago) period[14,15] and it is unclear whether the valley network formation or the TPW event caused by the Tharsis rise occurred first. To investigate whether the Tharsis bulge controlled the direction of the valley network, we modelled the stream network in a latitudinal band (from 40° S to the dichotomy) in the pre-TPW reference frame for a topography without Tharsis and before TPW at 1° per pixel (available in the Supplementary Information; see also Methods) and for a present-day topography with Tharsis at the same resolution. The directions of large-scale valley networks in the pre-TPW and in the present configuration are both compatible with the observed directions of valley network (Extended Data Figs 1 and 2). In both configurations valley networks are oriented towards the north, reflecting the topographic dichotomy of the Martian surface. This result indicates that the presence of the Tharsis bulge is not necessary to explain their orientation. However, the occurrence of valley networks within a palaeo tropical band, between the equator and 40° S, subject to precipitation (ice, snow or rainfall), supports a post-incision Tharsis-driven TPW during the Early/Late Hesperian period.

Early Mars climate simulations[16–18] assuming the present topography predict patchy ice/water accumulation in a south tropical band for a cold/icy scenario[17,18] and precipitation around Hellas basin and in Tharsis region for warm/wet conditions[18]. Both kinds of simulation fail to explain the occurrence of valley networks down to 45° S at the east of Hellas and down to 60° S at the south of Tharsis. They also predict substantial rain/snowfall in the west of Tharsis. This prediction does not match the lack of valley networks in this region. We performed new simulations using the same model and conditions (see Methods) using a pre-TPW topography. We found that ice tends to stabilize and accumulate in a tropical band (Fig. 2) similar to the distribution of valley networks (Fig. 1b), as a result of enhanced precipitation induced by adiabatic cooling when the atmospheric circulation transports water vapour southward, up to the highlands. An icy patch is also predicted in the Tharsis region but recent volcanic flows preclude the preservation of ancient morphologies in this region. Intensity of drainage is affected by several factors and is not only related to the intensity of precipitation. The geological history subsequent to valley network formation may also be responsible for heterogeneous modifications of the palaeo drainage intensities.

Can we find any geological clues of a past polar climate at the location of the palaeo poles? The palaeo north pole is located in Scandia Colles (Extended Data Fig. 4), a knobby terrain that probably represents one of the rare Noachian units in the north polar region[19]. Interestingly, some of the landforms of this region have been attributed to Late Hesperian polar ice retreat or melting[20,21], which could be the result of the

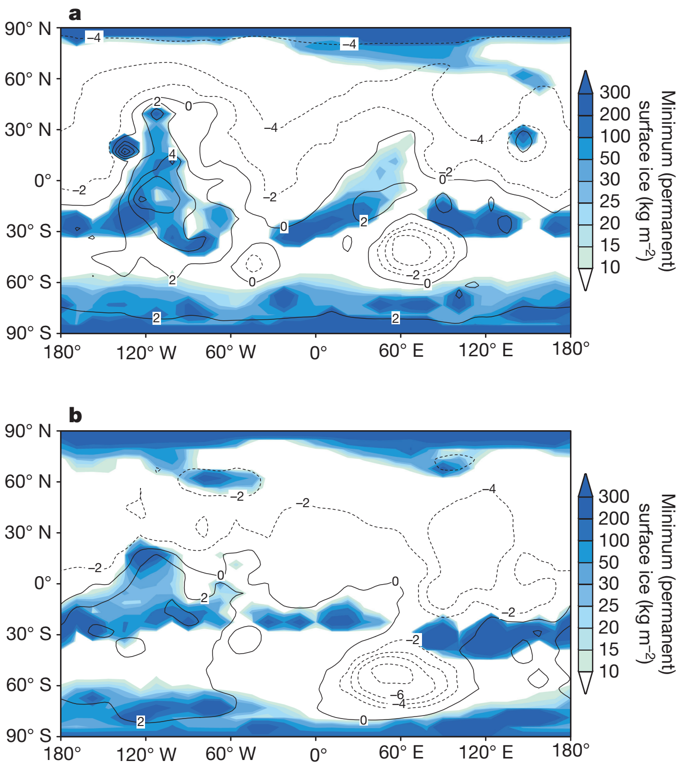


Figure 2 | Permanent ice deposits predicted by the global climate model for early Mars, with obliquity 45°, a circular orbit and mean surface pressure ∼0.2 bar. a, With present-day topography. **b**, With topography before TPW, without the Tharsis bulge. These maps show the yearly minimum of surface ice in simulations with just enough ice in the system to simulate the location where the ice stabilizes under a specific configuration. Elevations are given in kilometres, with solid lines representing isoaltitudes equal to or higher than 0 km and dashed lines representing isoaltitudes lower than 0 km.

displacement of the pole at this time. This palaeo north pole is located within the unique region of the northern plains where the SHARAD (Shallow Subsurface Radar) instrument has detected the base of ground ice at a depth of several tens of metres[22] that could correspond to the remnant of an ancient ice cap. Similarly, the southern hemisphere radar interface detected in the Dorsa Argentea region has been attributed to the remnant of an Hesperian-age buried ice cap[23]. The palaeo north pole is also surrounded by an area of anomalous neutron counting rates[24] associated with high hydrogen content (indicating water) in the first metre below the surface (Extended Data Fig. 5). Although the formation of this reservoir is likely to be recent, it may be related to the slow sublimation of deeper and older ice reservoirs favouring the stability of ice near the surface.

The palaeo south pole is located in the Noachian volcanic unit, Malea Planum. Some deposits in this region (Pityusa Patera) are possible remnants of an extensive polar ice sheet emplaced in the late Noachian to mid-Hesperian[25]. Erosional lineated terrains[26] and faintly dentritic Hesperian valley networks (Axius Valles and Mad Vallis)[25] are other features expected in a context of erosion by meltwater associated with retreating ice sheets[25,27]. A higher concentration of double-layered ejecta in this unit[28] also suggests the presence of an ice-rich substratum[28,29] that might be related to the polar palaeo location of this unit.

In summary, we consider that the planet was initially oriented such that the dichotomy was parallel to the equator (Fig. 3a) during the Noachian and the Early Hesperian periods[6]. Precipitation occurred in a tropical band during these periods. The incision of valley networks was contemporaneous with an initial phase of magmato-tectonic activity at Tharsis (Fig. 3b). The prolonged magmatic activity during the Hesperian moved Tharsis close to the present equator after the end of most of the fluvial activity (Fig. 3c). This TPW event may have generated stresses that were large enough to produce a global tectonic pattern (Extended Data Fig. 6). Global stress and tectonic pattern calculations show that extensional features oriented roughly north–south should be observed around the palaeo poles[7,30]. These fault patterns are not observed[7]. However, the identification of the expected global tectonic pattern may be complicated by additional stresses generated by Tharsis loading or the possibility of stress relaxation[30]. Stress relaxation may arise as a result of the expected slow TPW speed of about 1° per million years[31,32].

Causal links between volcanic outgassing, build-up of atmospheric pressure and precipitation have already been considered, but the valley networks were thought to have formed after the emplacement of most of the Tharsis load[33]. In light of our results, we conclude that precipitation occurred during the birth and growth of the Tharsis bulge rather than when its activity was declining. Such a scenario is more plausible than the post-Tharsis formation of valley networks, considering

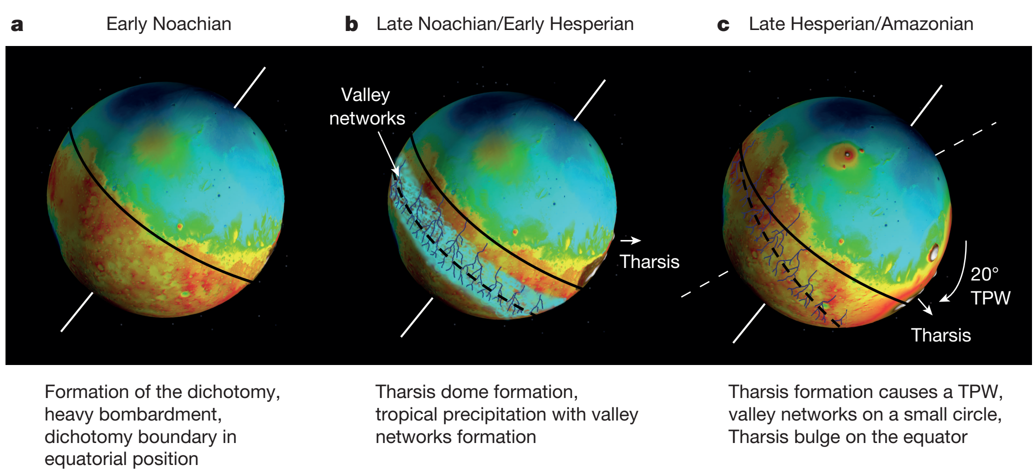


Figure 3 | Scenario for a TPW driven by a late growth of Tharsis contemporaneous with valley network incision. a, During the Early Noachian period, the planet was initially oriented such that the dichotomy was parallel to the equator[6]. **b**, Precipitation occurred in a tropical band during the Noachian and the Early Hesperian periods and was contemporaneous with the growth of Tharsis. **c**, The prolonged magmatic activity at Tharsis during the Hesperian period drove a slow TPW and moved Tharsis close to the present equator after the end of most of the fluvial activity.

early degassing and progressive depletion of volatiles in the mantle source. The calculated pre-Tharsis topographic map of Mars provides a framework within which to examine the first billion years of its geological history.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 July 2015; accepted 27 January 2016.
Published online 2 March 2016.

1. Phillips, R. J. *et al.* Ancient geodynamics and global-scale hydrology on mars. *Science* **291,** 2587–2591 (2001).
2. Anderson, R. C. *et al.* Primary centers and secondary concentrations of tectonic activity through time in the western hemisphere of Mars. *J. Geophys. Res.* **106,** 20563–20586 (2001).
3. Grott, M. *et al.* Long-term evolution of the martian crust-mantle system. *Space Sci. Rev.* **174,** 49–111 (2013).
4. Grott, M. Late crustal growth on Mars: evidence from lithospheric extension. *Geophys. Res. Lett.* **32,** L23201 (2005).
5. Nahm, A. L. & Schultz, R. A. Evaluation of the orogenic belt hypothesis for the formation of the Thaumasia highlands, Mars. *J. Geophys. Res.* **115,** E04008 (2010).
6. Roberts, J. H. & Zhong, S. The cause for the north south orientation of the crustal dichotomy and the equatorial location of Tharsis on Mars. *Icarus* **190,** 24–31 (2007).
7. Melosh, H. J. Tectonic patterns on a reoriented planet: Mars. *Icarus* **44,** 745–751 (1980).
8. Willemann, R. J. Reorientation of planets with elastic lithospheres. *Icarus* **60,** 701–709 (1984).
9. Rouby, H., Greff-Lefftz, M. & Besse, J. Rotational bulge and one plume convection pattern: influence on Martian true polar wander. *Earth Planet. Sci. Lett.* **272,** 212–220 (2008).
10. Matsuyama, I. & Manga, M. Mars without the equilibrium rotational figure, Tharsis, and the remnant rotational figure. *J. Geophys. Res.* **115,** E12020 (2010).
11. Carr, M. H. The Martian drainage system and the origin of valley networks and fretted channels. *J. Geophys. Res.* **100,** 7479–7507 (1995).
12. Hynek, B. M., Beach, M. & Hoke, M. R. T. Updated global map of Martian valley networks and implications for climate and hydrologic processes. *J. Geophys. Res.* **115,** E09008 (2010).
13. Irwin, R. P., III, Craddock, R. A., Howard, A. D. & Flemming, H. L. Topographic influences on development of Martian valley networks. *J. Geophys. Res.* **116,** E02005 (2011).
14. Fassett, C. I. & Head, J. W. III. The timing of Martian valley network activity: constraints from buffered crater counting. *Icarus* **195,** 61–89 (2008).
15. Bouley, S. & Craddock, R. A. Age dates of valley network drainage basins and subbasins within Sabae and Arabia Terrae, Mars. *J. Geophys. Res.* **119,** 1302–1310 (2014).
16. Forget, F. *et al.* 3D modelling of the early Martian climate under a denser CO_2 atmosphere: temperatures and CO_2 ice clouds. *Icarus* **222,** 81–99 (2013).
17. Wordsworth, R. *et al.* Global modelling of the early Martian climate under a denser CO_2 atmosphere: water cycle and ice evolution. *Icarus* **222,** 1–19 (2013).
18. Wordsworth, R. D. *et al.* Comparison of “warm and wet” and “cold and icy” scenarios for early Mars in a 3D climate model. *J. Geophys. Res.* **120,** 1201–1219 (2015).
19. Tanaka, K. & Kolb, E. Geologic history of the polar regions of Mars based on Mars Global Surveyor data. I. Noachian and Hesperian Periods. *Icarus* **154,** 3–21 (2001).
20. Fishbaugh, K. & Head, J. North polar region of Mars: topography of circumpolar deposits from Mars Orbiter Laser Altimeter (MOLA) data and evidence for asymmetric retreat of the polar cap. *J. Geophys. Res.* **105,** 22455–22486 (2000).
21. Tanaka, K. L. *et al.* History of plains resurfacing in the Scandia region of Mars. *Planet. Space Sci.* **59,** 1128–1142 (2011).
22. Putzig, N. E. *et al.* SHARAD soundings and surface roughness at past, present, and proposed landing sites on Mars: reflections at Phoenix may be attributable to deep ground ice. *J. Geophys. Res.* **119,** 1936–1949 (2014).
23. Kress, A. M. & Head, J. W. Late Noachian and early Hesperian ridge systems in the south circumpolar Dorsa Argentea Formation, Mars: evidence for two stages of melting of an extensive late Noachian ice sheet. *Planet. Space Sci.* **109–110,** 1–20 (2015).
24. Feldman, W. C. *et al.* Global distribution of near-surface hydrogen on Mars. *J. Geophys. Res.* **109,** E09006 (2004).
25. Head, J. W. & Pratt, S. Extensive Hesperian-aged south polar ice sheet on Mars: evidence for massive melting and retreat, and lateral flow and ponding of meltwater. *J. Geophys. Res. Planets* **106,** 12275–12299 (2001).
26. Kargel, J. S. & Strom, R. G. Ancient glaciation on Mars. *Geology* **20,** 3–7 (1992).
27. Leonard, G. J. & Tanaka, K. L. Geologic map of the Hellas region of Mars. *USGS Surv. Misc. Invest. Ser.* Map I–2694 (scale 1:4,336,000) http://pubs.usgs.gov/imap/i2694/ (USGS, 2001).
28. Costard, F. The spatial distribution of volatiles in the martian hydrolithosphere. *Earth Moon Planets* **45,** 265–290 (1989).
29. Weiss, D. K. & Head, J. W. Formation of double-layered ejecta craters on Mars: a glacial substrate model. *Geophys. Res. Lett.* **40,** 3819–3824 (2013).
30. Grimm, R. E. & Solomon, S. C. Tectonic tests of proposed polar wander paths for Mars and the Moon. *Icarus* **65,** 110–121 (1986).
31. Tsai, V. C. & Stevenson, D. J. Theoretical constraints on true polar wander. *J. Geophys. Res.* **112,** B05415 (2007).
32. Chan, N. H. *et al.* Time-dependent rotational stability of dynamic planets with elastic lithospheres. *J. Geophys. Res.* **119,** 169–188 (2014).
33. Bibring, J. P. *et al.* Global mineralogical and aqueous Mars history derived from OMEGA/Mars Express data. *Science* **312,** 400–404 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was funded by the GEOPS laboratory, the Programme National de Planétologie of INSU-CNRS and the Centre National d’Etude Spatiale (CNES).

Author Contributions S.B. conceived the project. S.B and D.B. drafted the manuscript with contributions from all authors and performed calculations of palaeo poles from valley networks distribution. I.M. performed the calculation of the rotational figure of Mars and its surface topography before TPW and Tharsis. F.F. and M.T. performed early Mars climate model simulations applied to the pre-TPW topography. A.S. and S.B. performed calculations of stream network for a topography of Mars with and without Tharsis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.B. (sylvain.bouley@u-psud.fr).

METHODS

Topography model. The topography without Tharsis and rotational deformation due to TPW was calculated using gravity and topography data[10]. We adopt the expected rotation pole location before the emplacement of Tharsis (259.5° E, 71.1° N), the expected elastic lithosphere thickness at the time of loading (58 km), and the expected Tharsis gravity and shape coefficients (tables 3–5 in ref. 10).

The gravitational potential external to Mars, at a point with spherical coordinates (r, θ, ϕ), where θ is co-latitude and ϕ is longitude, can be expanded in spherical harmonics with unnormalized expansion coefficients C_{lm} and S_{lm} as follows (see, for example, refs 34 and 35):

$$U(r,\theta,\phi)=\frac{GM}{r}+\frac{GM}{r}\sum_{l=2}^{\infty}\sum_{m=0}^{l}\left(\frac{R}{r}\right)^{l}\times P_{lm}(\cos\theta)[C_{lm}\cos(m\phi)+S_{lm}\cos(m\phi)] \quad (1)$$

where G is the gravitational constant, M and R are the mass and mean radius of Mars, and P_{lm} are unnormalized associated Legendre functions given by:

$$P_{lm}(\mu)=(1-\mu^2)^{m/2}\frac{\mathrm{d}^m}{\mathrm{d}\mu^m}P_l(\mu) \quad (2)$$

where P_l are Legendre polynomials[35,36]. We do not include the Condon–Shortey phase factor of $(-1)^m$. The unnormalized coefficients are related to the normalized expansion coefficients, $\bar{C}_{lm}$ and $\bar{S}_{lm}$:

$$\begin{pmatrix}C_{lm}\\S_{lm}\end{pmatrix}=\left[(2-\delta_{m0})(2l+1)\frac{(l-m)!}{(l+m)!}\right]^{1/2}\begin{pmatrix}\bar{C}_{lm}\\\bar{S}_{lm}\end{pmatrix} \quad (3)$$

where δ_{m0} is the Kronecker delta function.

We expand the geoid, N, and shape, S, in spherical harmonics with unnormalized expansion coefficients as:

$$N=R\sum_{l=2}^{\infty}\sum_{m=0}^{l}P_{lm}(\cos\theta)[C_{lm}\cos(m\phi)+S_{lm}\sin(m\phi)]-\frac{1}{2}\frac{w^2R^4}{GM}\cos^2\theta-R\sum_{l=2}^{\infty}P_{l0}(\pi/2)C_{l0} \quad (4)$$

where the equipotential is chosen to be the mean potential at the equator, and:

$$S=\sum_{l=2}^{\infty}\sum_{m=0}^{l}P_{lm}(\cos\theta)[c_{lm}\cos(m\phi)+s_{lm}\sin(m\phi)] \quad (5)$$

where c_{lm} and s_{lm} are the shape expansion coefficients. We use the Jet Propulsion Laboratory Mars gravity field MRO95A[37] and the Mars Orbiting Laser Altimetry (MOLA) shape model[38]. The topography is given by $T\equiv S-N$.

We remove rotational deformation contributions associated with the change in the rotational potential due to TPW. Given the location of the rotation pole, (θ_R, ϕ_R), the rotational potential expansion coefficients are:

$$\begin{bmatrix}C_{2m}^{R}(\theta_R,\phi_R)\\S_{2m}^{R}(\theta_R,\phi_R)\end{bmatrix}=-(2-\delta_{m0})\frac{(2-m)!}{(2+m)!}\times\frac{1}{3}\times\frac{\Omega^2R^3}{GM}P_{2m}(\cos\theta_R)\begin{bmatrix}\cos(m\phi_R)\\\sin(m\phi_R)\end{bmatrix} \quad (6)$$

where Ω is the rotation rate. The gravity expansion coefficients associated with rotational deformation due to TPW can be written as:

$$\begin{bmatrix}C_{2m}^{TPW}\\S_{2m}^{TPW}\end{bmatrix}=k_2\begin{bmatrix}C_{2m}^{R}(\theta_{R,f},\phi_{R,f})\\S_{2m}^{R}(\theta_{R,f},\phi_{R,f})\end{bmatrix}-k_2\begin{bmatrix}C_{2m}^{R}(\theta_{R,i},\phi_{R,i})\\S_{2m}^{R}(\theta_{R,i},\phi_{R,i})\end{bmatrix} \quad (7)$$

where k_2 is the degree-2 tidal Love number describing the long-term gravity deformation due to the change in the rotational potential, and $(\theta_{R,f}, \phi_{R,f})$ and $(\theta_{R,i}, \phi_{R,i})$ are the final (present) and initial (before TPW) spherical coordinates of the rotation pole respectively. Similarly, the shape expansion coefficients associated with rotational deformation due to TPW can be written as:

$$\begin{bmatrix}c_{2m}^{TPW}\\s_{2m}^{TPW}\end{bmatrix}=Rh_2\begin{bmatrix}C_{2m}^{R}(\theta_{R,f},\phi_{R,f})\\S_{2m}^{R}(\theta_{R,f},\phi_{R,f})\end{bmatrix}-Rh_2\begin{bmatrix}C_{2m}^{R}(\theta_{R,i},\phi_{R,i})\\S_{2m}^{R}(\theta_{R,i},\phi_{R,i})\end{bmatrix} \quad (8)$$

where h_2 is the degree-2 tidal displacement Love number describing the long-term shape deformation due to the change in the rotational potential.

The dimensionless k_2 and h_2 Love numbers depend on Mars' interior structure and rheology. We assume the five-layer internal structure model described in table 2 of ref. 10, and use the classical propagator matrix method (for example, ref. 39) to calculate their values. For the expected elastic lithosphere thickness of 58 km, $k_2=1.10$ and $h_2=2.00$. These Love numbers are not sensitive to elastic lithosphere thickness. For example, $k_2=1.19$ and $h_2=2.19$ for a model without an elastic lithosphere.

We compute the expansion coefficients for the topography without Tharsis and the rotational deformation associated with TPW by removing these contributions from the observed gravity and shape coefficients. The Tharsis gravity and shape coefficients are taken from tables 3–5 of ref. 10, and the gravity and shape coefficients for the rotational deformation associated with TPW are given by equations (6) and (7). Finally, we compute the spherical harmonic coefficients in the pre-TPW frame using Wigner-D functions.

Determination of the palaeo poles. The palaeo pole positions are calculated from a least-squares adjustment of the valley network density to a small circle on the sphere. (A small circle is given by the intersection of the sphere with a plane.) Considering the plane equation in Cartesian coordinates:

$$z=ax+by+c \quad (9)$$

and the conversion equation of spherical coordinates (r, λ, φ) into Cartesian coordinates, where r is the average radius of Mars, λ is the longitude, and φ the latitude:

$$\begin{aligned}x&=r\cos\lambda\cos\varphi\\y&=r\sin\lambda\cos\varphi\\z&=r\sin\varphi\end{aligned} \quad (10)$$

the equation of a small circle is given by:

$$\tan\varphi_{sc}=a\cos\lambda_{sc}+b\sin\lambda_{sc}+\frac{c}{r\cos\varphi_{sc}} \quad (11)$$

This equation may be solved with $u=\cos(\varphi_{sc})$ and by extracting the roots of the second-degree polynomial in u. We thus have:

$$\varphi_{sc}=\arccos\left(\frac{-\frac{2A}{r}\pm\sqrt{\Delta}}{2[A^2+1]}\right) \quad (12)$$

where:

$$A=a\cos\lambda_{sc}+b\sin\lambda_{sc} \quad (13)$$

and:

$$\Delta=\left(\frac{2Ac}{r}\right)^2-4\left(\frac{c^2}{r^2}-1\right)(A^2+1) \quad (14)$$

The parameters a, b and c are determined by minimizing of the sum of the weighted residuals between calculated latitudes and observed latitudes of the valley network density map:

$$v=\sum_i[\varphi_{sc}(\lambda_i)-\varphi_i]^2d_i^2 \quad (15)$$

where λ_i and φ_i are coordinates of the valley network density map and d_i is the corresponding density of valley networks[12]. Adjustment is achieved by direct exploration of the three-dimensional parameter space. This approach is useful to determine the best solution corresponding to the minimum of v, but also to determine the spread of solutions (a, b, c) corresponding to any chosen residual greater than the minimum value. The residual may be expressed as a root mean square (r.m.s.) in latitude using:

$$\text{r.m.s.}=\sqrt{\frac{v_{min}}{\sum_i d_i^2}} \quad (16)$$

The palaeo pole positions are then given from parameters a and b and the following equations:

$$\begin{aligned}\varphi_{pole}&=90-\arccos\left(\frac{1}{\sqrt{a^2+b^2+1}}\right)\\\lambda_{pole}&=\arctan\left(\frac{b}{a}\right)\end{aligned} \quad (17)$$

The best solution is found at 24° S and corresponds to a r.m.s. in latitude of 14° that is equal to the spread of valley networks in a direction perpendicular to the small circle. The best palaeo pole positions are indicated with a diamond in Fig. 1a, together with the spread in longitude and latitude. Colours from black to red correspond to the associated r.m.s. value for each solution, as given by equation (16), and vary from 14° (black) to 15° (red).

Determination of the stream network before and after Tharsis emplacement. The relief controls the direction of runoff processes. To investigate whether the emplacement of the Tharsis bulge controlled the direction of the valley network, the stream network was modelled for a topography of Mars without Tharsis and for a topography with Tharsis with the same resolution of 1° per pixel. We used the Arc Hydro tool in ArcGIS (http://downloads.esri.com/archydro/archydro/) that includes different functions to extract hydrological parameters (that is, the flow direction, the flow accumulation and the stream definition[40]). The digital elevation model (DEM) can contain artefacts of DEM construction, but these have been corrected.

The flow direction is based exclusively on topography. For each cell, the flow direction corresponds to the direction of the steepest slope between the cell and the eight neighbouring cells. The result is a raster with the value of every cell corresponding to one of the eight possible flow directions. Then, based on the flow direction raster, for each cell, the flow accumulation is calculated as the total number of cells drained upstream of each cell. The flow accumulation approximately represents the drainage network.

Finally, the stream network is defined, based on the flow accumulation network and on a user-defined river threshold value (Extended Data Fig. 2). For each cell, if the value is greater than this threshold, it is defined as a stream. A smaller threshold will result in a denser stream network and usually in a greater number of delineated catchments, which may hinder delineation performance. For both DEMs, in order to achieve general flow directions, which are not meant to correspond directly to the actual valley network, we chose a river threshold value of 15 cells.

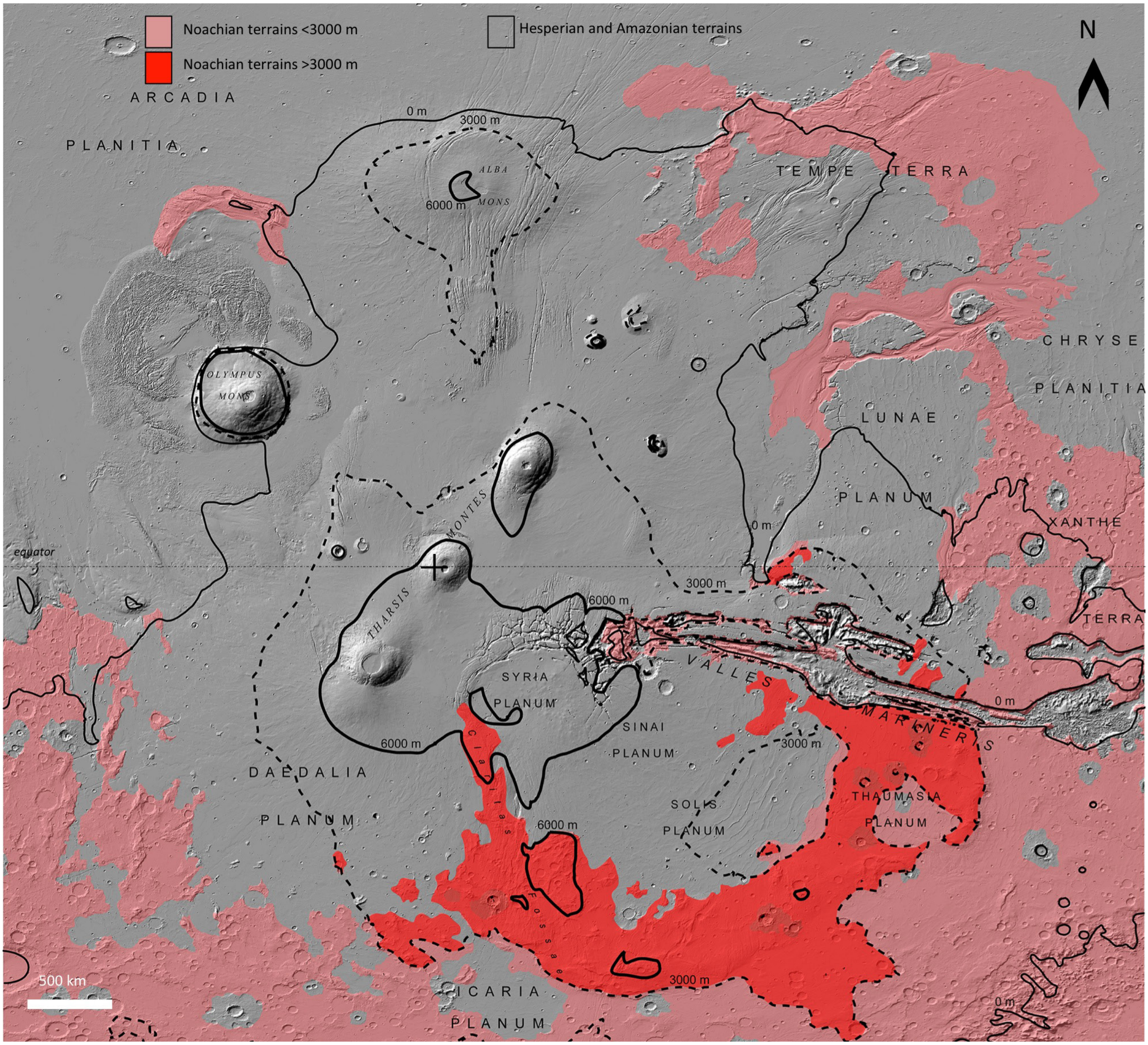
The orientation of the stream network was calculated by using an ArcGIS tool[41]. This orientation was calculated with respect to the north (azimuth). The data were compiled in a rose diagram representing the orientation of the stream network for the DEM without Tharsis (total number of measurements $N=702$; Extended Data Fig. 3a) and with Tharsis ($N=698$; Extended Data Fig. 3b).

The distributions of the stream network for both topographies (before and after Tharsis emplacement and TPW) are similar (Extended Data Fig. 2). The stream network is mainly oriented towards the north in both configurations (Extended Data Fig. 3). The calculation shows that the flow direction in the highlands was primarily controlled by the dichotomy and limits the possible influence of the Tharsis bulge.

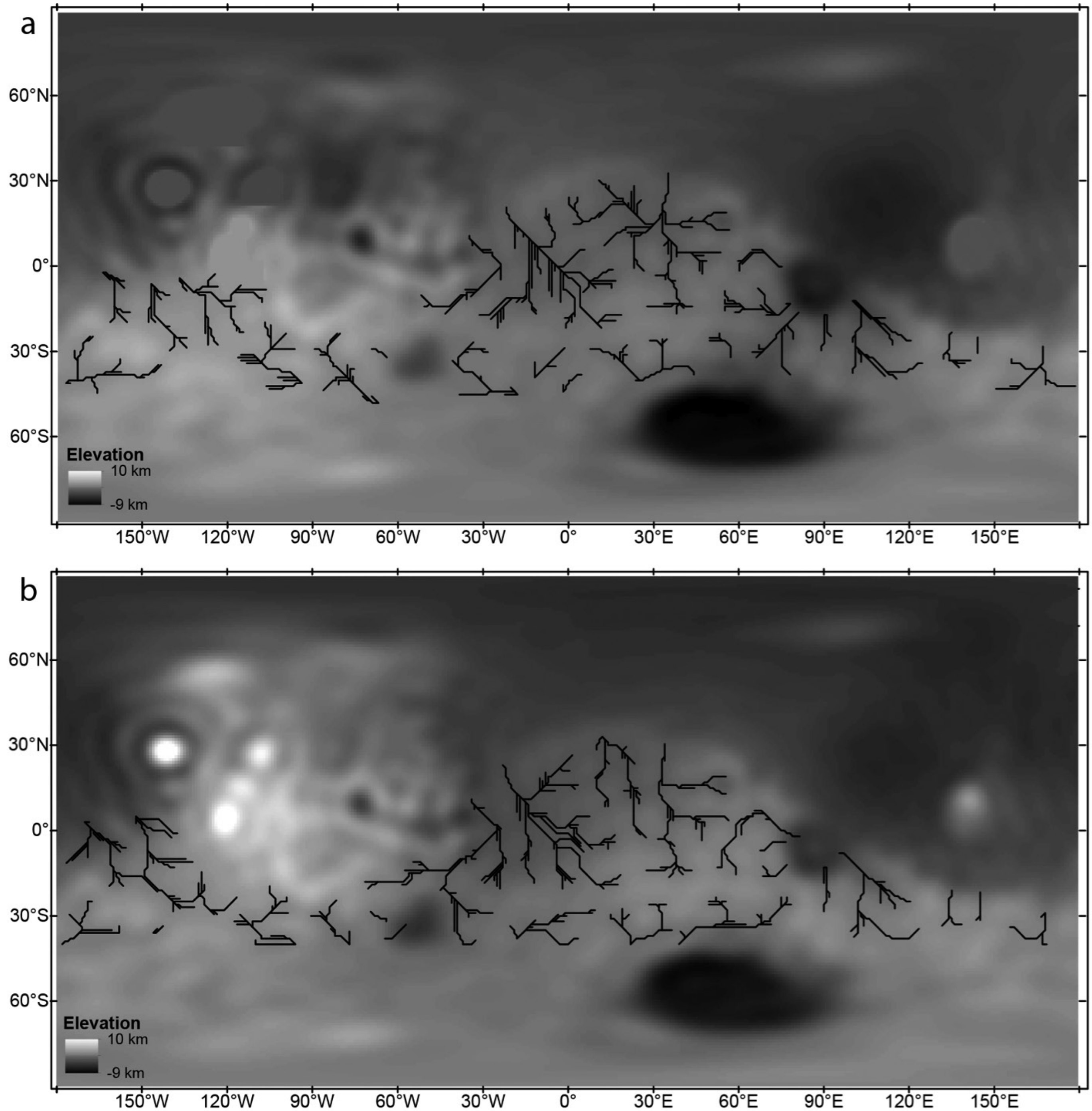
Global climate model simulations. We use the Laboratoire de Météorologie Dynamique Early Mars 3D global climate model[16,17]. It includes the modelling of the CO_2 cycle (condensation and sublimation on the surface and in the atmosphere), a water cycle (transport of water vapour and clouds, precipitation and evaporation) and a detailed radiative transfer code adapted to a thick CO_2 atmosphere and both CO_2 and H_2O clouds. Here we used the ice equilibration algorithm from ref. 17 designed to calculate the location where the ice deposits stabilize at equilibrium under early Mars conditions, after the equivalent of thousands of years of evolution. This algorithm was shown to be insensitive of the assumed initial state and in particular to the location of the ice reservoir at the beginning of the simulation.

Figure 2 presents a map of the permanent ice deposits predicted for the most likely conditions for early Mars, that is, with an obliquity of 45° (ref. 42) and a mean pressure of 0.2 bar of CO_2 (ref. 16). Similar results were obtained with a mean pressure of 1 bar.

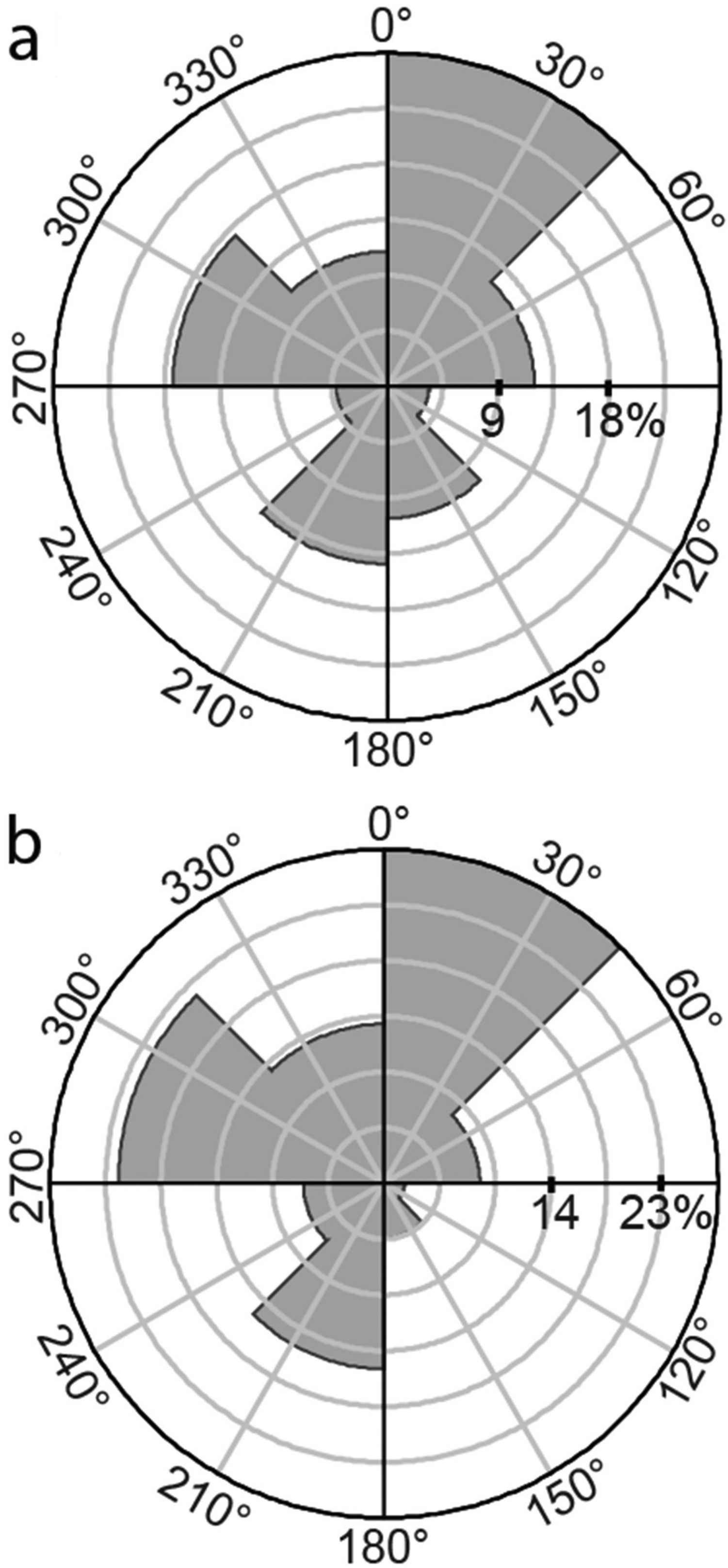
34. Kaula, W. M. *An Introduction to Planetary Physics: the Terrestrial Planets* (John Wiley & Sons, 1968).
35. Wieczorek, M. A. in *Treatise on Geophysics* 165–206 (2007).
36. Arfken, G. & Weber, H. *Mathematical Methods for Physicists* 4th edn (Academic Press, 1995).
37. Lemoine, F. G., Konopliv, M. & Zuber, M. T. MRO Derived Gravity Science Data Products, MRO-M-RSS-5-SDP-V1.0, NASA Planetary Data System, https://pds.nasa.gov/ds-view/pds/viewProfile.jsp?dsid=MRO-M-RSS-5-SDP-V1.0 (2008).
38. Smith, D. E. MOLA initial experiment gridded data record, MGS-M-MOLA-5-IEGDR-L3-V1.0, NASA Planetary Data System, https://pds.nasa.gov/ds-view/pds/viewDataset.jsp?dsid=MGS-M-MOLA-5-IEGDR-L3-V1.0 (1999).
39. Sabadini, R. & Vermeersen, B. *Global Dynamics of the Earth: Applications of Normal Mode Relaxation Theory to Solid-Earth Geophysics* (Kluwer Academic, 2004).
40. ESRI. *Arc Hydro Tools Overview* http://downloads.esri.com/blogs/hydro/ah2/arc_hydro_tools_2_0_overview.pdf (Environmental Systems Research Institute, 2004).
41. Jenness, J. S. Some thoughts on analyzing topographic habitat characteristics. In *Remotely Wild* http://www.jennessent.com/downloads/topographic_analysis_online.pdf (GIS, Remote Sensing, and Telemetry Working Group of The Wildlife Society, June 2005).
42. Laskar, J. *et al.* Long term evolution and chaotic diffusion of the insolation quantities of Mars. *Icarus* **170,** 343–364 (2004).
43. Tanaka, K. L. *et al. Geologic map of Mars: U.S. Geological Survey Scientific Investigations Map 3292, scale 1:20,000,000* http://dx.doi.org/10.3133/sim3292 (2014).



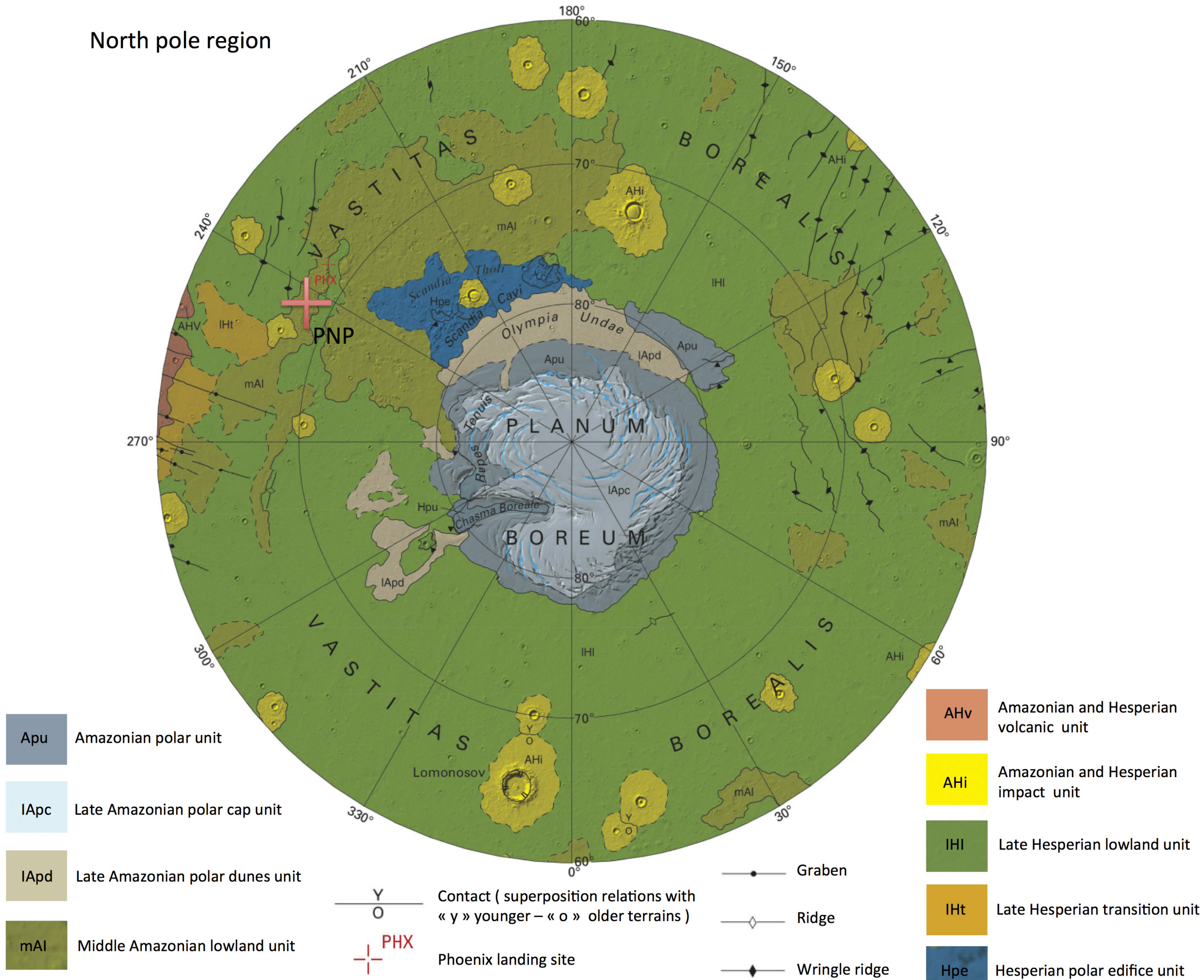
Extended Data Figure 1 | Map of Tharsis region with 0 m, 3,000 m and 6,000 m isoaltitude lines. Noachian terrains are mapped in light red for terrains lower than 3,000 m and in dark red for terrains higher than 3,000 m. Hesperian and Amazonian terrains are in grey. Age units are taken from the most recent geological map of this region[43]. The black cross on Tharsis Montes is the location of the centre of mass of the Tharsis dome.



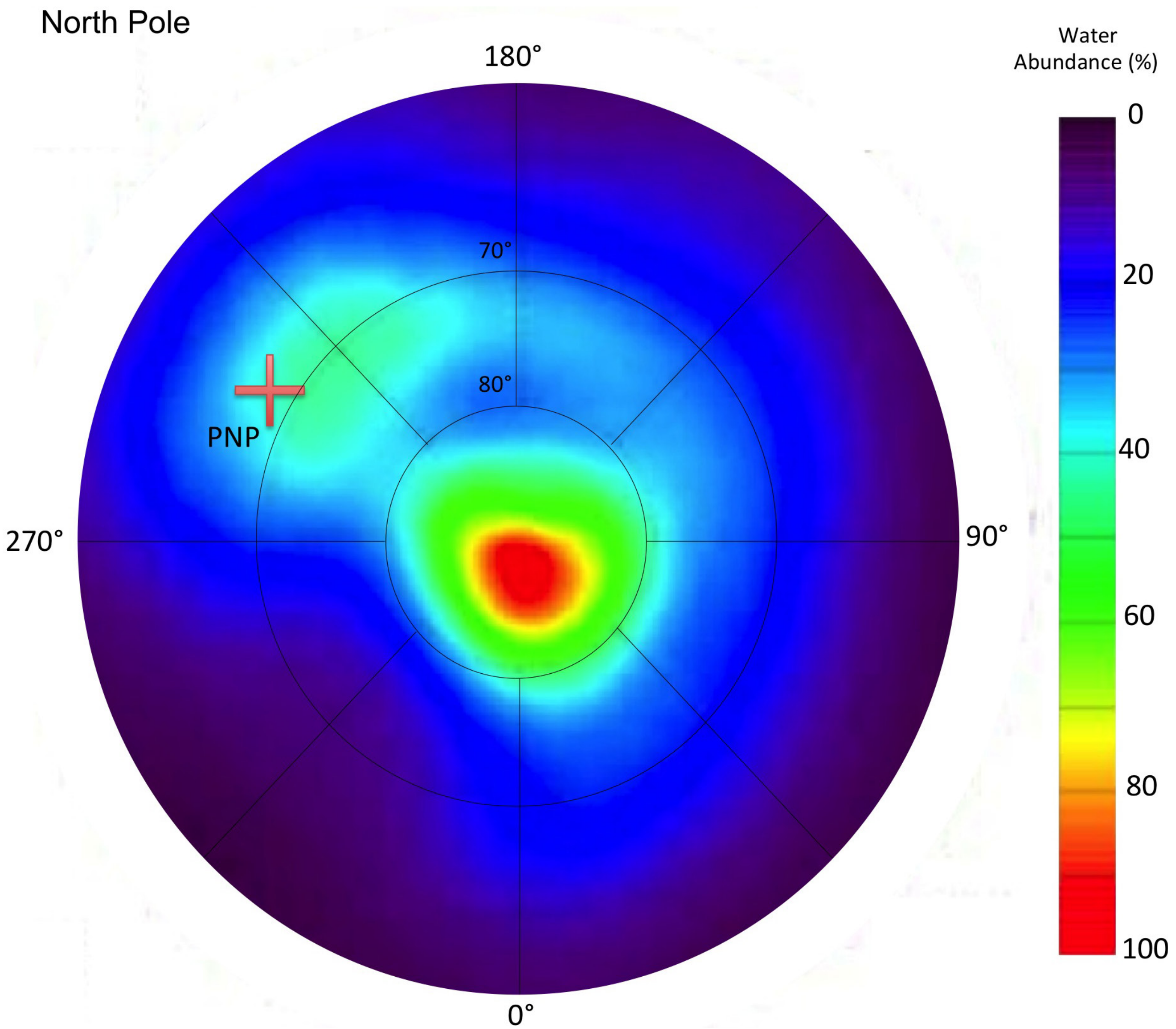
Extended Data Figure 2 | Modelled stream network before and after Tharsis emplacement. a, **b**, Digital Elevation Model (DEM) with 1° per pixel resolution without Tharsis (**a**) and with Tharsis (**b**). The stream network was modelled using the Arc Hydro tool in ArcGIS.



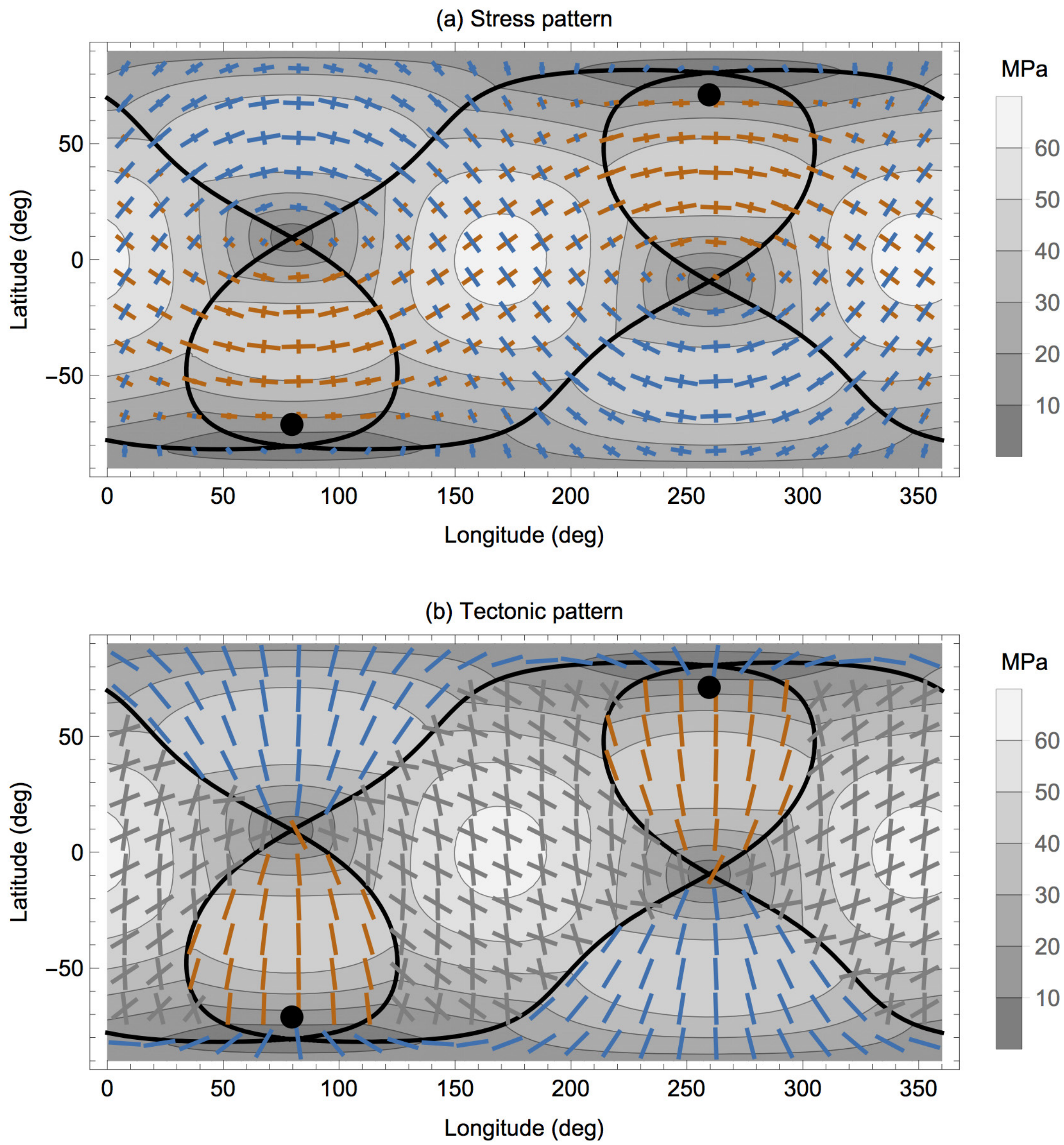
Extended Data Figure 3 | Rose diagram of orientations of the modelled stream network. a, Before Tharsis emplacement ($N = 702$). **b**, After Tharsis emplacement ($N = 698$). The orientation values are grouped into 45° sectors. N is the total number of orientation measurements.



Extended Data Figure 4 | Geological map of the north polar region. The red cross indicates the location of the palaeo north pole (PNP), inferred from the valley network distribution. Figure modified from ref. 43; US Geological Survey.



Extended Data Figure 5 | Orthographic projection of lower-limit concentrations of water abundance at latitudes poleward of 50° N. The red cross indicates the location of the palaeo north pole, inferred from the valley network distribution. Figure modified with permission from figure 5 of Feldman, W. C. *et al.*[24], *J. Geophys. Res.*, John Wiley and Sons, copyright 2004 by the American Geophysical Union.



Extended Data Figure 6 | Predicted global-scale stress and tectonic patterns due to the Tharsis-driven TPW event. Solid circles indicate the locations of the palaeo poles. In the stress pattern (**a**), crosses indicate directions and relative magnitudes of principal stresses, and orange and blue lines correspond to extensional and compressive stresses, respectively. In the tectonic pattern (**b**), the orange, blue, and light grey lines indicate the strike of the expected normal, thrust and strike–slip faults, respectively. Contours correspond to the deviator stress in units of MPa. Solid black lines mark the boundaries between different tectonic regions.

doi:10.1038/nature16984

Enhancing coherence in molecular spin qubits via atomic clock transitions

Muhandis Shiddiq[1]*, Dorsa Komijani[1]*, Yan Duan[2], Alejandro Gaita-Ariño[2], Eugenio Coronado[2] & Stephen Hill[1]

Quantum computing is an emerging area within the information sciences revolving around the concept of quantum bits (qubits). A major obstacle is the extreme fragility of these qubits due to interactions with their environment that destroy their quantumness. This phenomenon, known as decoherence, is of fundamental interest[1,2]. There are many competing candidates for qubits, including superconducting circuits[3], quantum optical cavities[4], ultracold atoms[5] and spin qubits[6–8], and each has its strengths and weaknesses. When dealing with spin qubits, the strongest source of decoherence is the magnetic dipolar interaction[9]. To minimize it, spins are typically diluted in a diamagnetic matrix. For example, this dilution can be taken to the extreme of a single phosphorus atom in silicon[6], whereas in molecular matrices a typical ratio is one magnetic molecule per 10,000 matrix molecules[10]. However, there is a fundamental contradiction between reducing decoherence by dilution and allowing quantum operations via the interaction between spin qubits. To resolve this contradiction, the design and engineering of quantum hardware can benefit from a 'bottom-up' approach whereby the electronic structure of magnetic molecules is chemically tailored to give the desired physical behaviour. Here we present a way of enhancing coherence in solid-state molecular spin qubits without resorting to extreme dilution. It is based on the design of molecular structures with crystal field ground states possessing large tunnelling gaps that give rise to optimal operating points, or atomic clock transitions, at which the quantum spin dynamics become protected against dipolar decoherence. This approach is illustrated with a holmium molecular nanomagnet in which long coherence times (up to 8.4 microseconds at 5 kelvin) are obtained at unusually high concentrations. This finding opens new avenues for quantum computing based on molecular spin qubits.

One of the proposed approaches to obtaining spin qubits is that of using magnetic molecules[8–16]. Up to now, coherence has been optimized through dilution and deuteration to minimize dipolar and hyperfine interactions, respectively[9–11,13,16]. A class of molecules in which these two sources of decoherence can be minimized by alternative means are the so-called polyoxometalates. In the past, these metal oxide clusters have been used as model systems in molecular magnetism because of their ability to host magnetic ions in chemically tailored environments of high symmetry and rigidity[17]. Currently, these molecules are seen as potential building blocks in quantum computing architectures[18–22].

In the present study we chose the $[Ho(W_5O_{18})_2]^{9-}$ complex (abbreviated HoW_{10}), which has been subjected to extensive structural, magnetic and spectroscopic characterizations that raised the possibility of observing coherent spin dynamics[23,24]. HoW_{10} is formed by two molecular tungsten oxide moieties encapsulating a Ho^{3+} ion (Fig. 1). The geometry around Ho^{3+} exhibits a slightly distorted square-antiprismatic environment, which can be approximated by a D_{4d} 'pseudo-axial' symmetry. This results in a splitting of the total angular momentum, $J=8$, ground state spin–orbit manifold according to its m_J quantum numbers. Quantitatively this splitting can be described by the spin Hamiltonian in equation (1), where the double summation parameterizes the crystal-field (CF) interaction which, for D_{4d} symmetry, contains the axial terms $B_2^0\hat{O}_2^0$, $B_4^0\hat{O}_4^0$ and $B_6^0\hat{O}_6^0$ (see Methods for definition and discussion of terms in equation (1)[24])

$$\hat{H}=\sum_{k=2,4,6}\sum_{q=0}^{k}B_k^q\hat{O}_k^q+\hat{J}\cdot A\cdot\hat{I}+\mu_B\boldsymbol{B}_0\cdot g_L\cdot\hat{J}-\mu_N g_N\boldsymbol{B}_0\cdot\hat{I} \qquad (1)$$

This results in an isolated $m_J=\pm4$ ground doublet, separated from the first excited states ($m_J=\pm5$) by $\sim$20 cm^{-1}. This picture provides a reasonable description of the magnetic properties of this molecule[23]. However, minor deviations from D_{4d} symmetry that are present in the crystal make operative the tetragonal $B_4^4\hat{O}_4^4$ CF interaction. Interestingly, the match between the ($\pm$integer) values of the ground state spin projections, $m_J=\pm4$, and the tetragonal (that is, $q=4$) order of the main symmetry axis of the molecule results in the $B_4^4\hat{O}_4^4$ (here $\hat{O}_4^4=\frac{1}{2}(\hat{J}_+^4+\hat{J}_-^4)$) interaction generating an unusually large quantum tunnelling gap, $\Delta\approx9.18$ GHz ($\sim$0.3 cm^{-1})[24]. This gap is a crucial factor for the coherence of electron spin dynamics in molecular spin qubits, and is the main subject of the present study.

The standard approach to probing coherent spin dynamics involves the use of electron spin echoes in pulsed electron paramagnetic resonance (EPR). The HoW_{10} system is attractive in this regard because its predicted tunnelling gap ($\sim$9.18 GHz, Fig. 1) is close to the X-band frequency associated with the most sophisticated EPR spectrometers. While the magnitude of the gap is set by B_4^4, interesting details of the EPR spectra are also determined by the hyperfine interaction between the Ho electron and nuclear spins (the second term in equation (1)). Holmium occurs naturally in only one stable isotope (^{165}Ho) with a nuclear spin of $I=7/2$. A strong hyperfine coupling ($A_\parallel=830\pm10$ MHz) results in the observation of eight ($2I+1$) well-resolved transitions via continuous-wave (CW) high-frequency EPR measurements[24]. The energy level scheme that arises from the combination of CF and hyperfine coupling, together with the Zeeman interaction (third and fourth terms in equation (1)), gives rise to a series of avoided level crossings between $m_J=\pm4$ states (with the same m_I), resulting in multiple gaps in the energy diagram near zero field (Fig. 1a).

Single crystals of $Na_9[Ho_xY_{(1-x)}(W_5O_{18})_2]\cdot nH_2O$ (where Y (yttrium) is non-magnetic) were prepared with Ho concentrations ranging from $x=0.25$ to $x=0.001$, that is, up to three orders of magnitude away from the usual high-dilution limit[10], allowing a study of the effects of dilution on electron dipolar spin–spin decoherence. Figure 2a displays electron-spin-echo- (ESE-) detected EPR spectra recorded at 5 K for a dilute ($x=0.001$) sample at frequencies from 9.1 GHz to 9.8 GHz, with $\theta=29°$ (θ is the angle between the applied field, $\boldsymbol{B}_0$, and the z axis

[1]National High Magnetic Field Laboratory and Department of Physics, Florida State University, Tallahassee, Florida 32310, USA. [2]Instituto de Ciencia Molecular, Universidad de Valencia, C/Catedrático José Beltrán 2, 46980 Paterna, Spain.
*These authors contributed equally to this work.

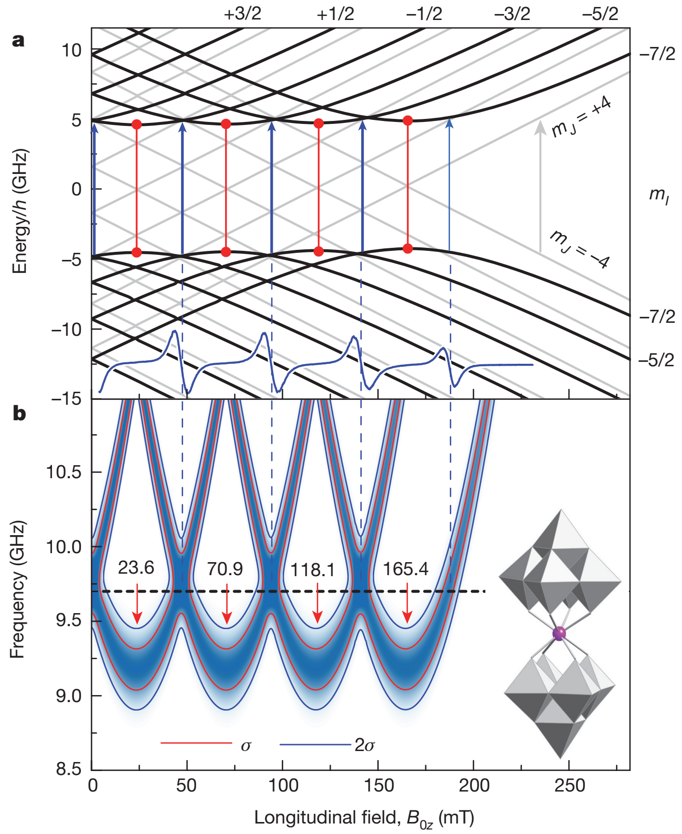


Figure 1 | HoW_{10} tunnelling gap. a, Main figure, Zeeman diagrams for the $m_J = \pm 4$, $I = 7/2$ ground state, with $\boldsymbol{B}_0 \| z$: the thin grey lines assume exact D_{4d} symmetry, while the thick black curves assume an 'axial + $B_4^4\hat{O}_4^4$' parameterization[24]. Numbers on the top and right axes indicate m_I values. Inset (blue trace at bottom), the corresponding 9.64 GHz CW EPR spectrum (reproduced from figure 8 of ref. 24 with permission from the Royal Society of Chemistry) is observed well below the D_{4d} prediction (grey arrow denotes expected highest field resonance), providing evidence for the tunnelling gap. Indeed, the 'axial + $B_4^4\hat{O}_4^4$' parameterization gives excellent agreement with the data, both in terms of resonance positions (blue arrows) and intensity (arrow thickness). The red vertical lines, meanwhile, indicate the locations of CTs. **b**, Main figure, 3D EPR intensity map including inhomogeneous broadening due to a Gaussian distribution in B_4^4 ($\sigma_{B44} = 2.1 \times 10^{-5}\ \mathrm{cm}^{-1}$); darker shading represents stronger intensity, with contours at the $\pm\sigma_\Delta$ (red) and $\pm 2\sigma_\Delta$ (blue) levels ($\sigma_\Delta = 123$ MHz, the s.d. in Δ). Red arrows denote CTs and dashed lines denote locations of 9.64 GHz resonances. Inset, the HoW_{10} molecule.

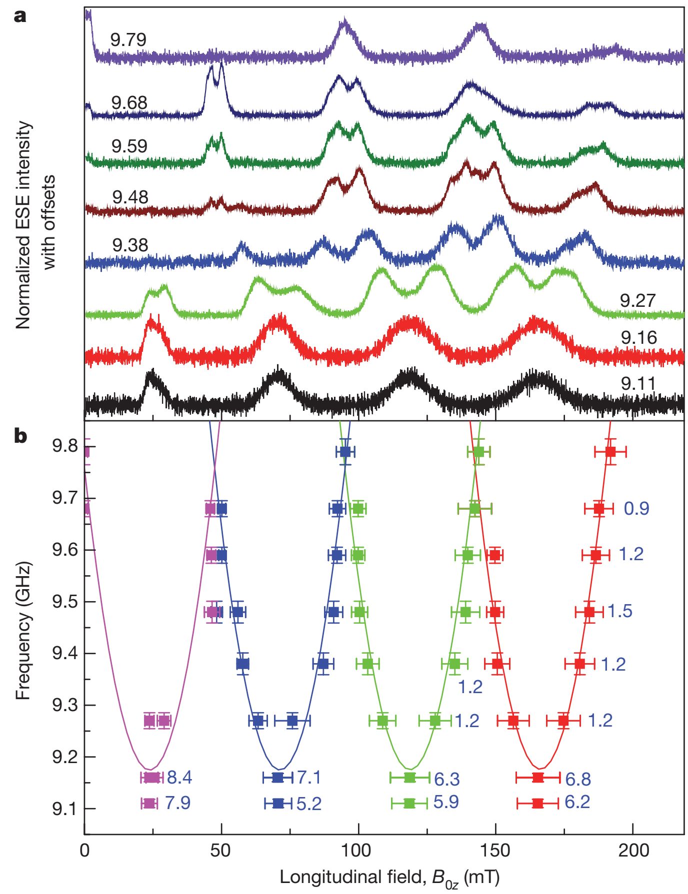


Figure 2 | ESE-detected spectra for a dilute sample. a, Variable frequency measurements at 5.0 K for an $x = 0.001$ crystal, with $\theta = 29°$; the frequencies are indicated in GHz above each trace. **b**, Frequency versus field plot of the resonances in **a**. The data are in good agreement with simulations (solid curves) based on the 'axial + $B_4^4\hat{O}_4^4$' parameterization[24]. Selected T_2 values (in μs) determined from the measurements in Fig. 3 are indicated in blue close to some of the data points. Vertical error bars in **b** denote pulse excitation bandwidths ($\pm 1/2\tau_{\pi/2}$, where $\tau_{\pi/2}$ is the duration of the π/2 pulse), while horizontal error bars represent standard deviations (±s.d.) deduced from Gaussian fits to the resonances in **a**.

of the crystal); ESE signals were generated using a two-pulse Hahn-echo sequence (see Methods)[25]. Four broad peaks of equal intensity are observed at the two lowest frequencies (9.11 GHz and 9.16 GHz), which were selected to be close to the gap minima in Fig. 1b. With increasing frequency, these peaks split and move symmetrically apart, as expected on the basis of predictions in Fig. 1b. For the most part, the data lie on the simulated curves, with the obvious exception of the two lowest frequencies and some lower-field (<60 mT) data points. The simulations are based on previously determined Hamiltonian parameters[24], and the spectra are plotted against the re-scaled longitudinal applied field, B_{0z} ($=B_0\cos\theta$), to facilitate comparisons between different samples (see Methods).

Two-pulse ESE measurements were separately used to determine 5 K transverse relaxation times, T_2, at selected points within the spectrum for the $x = 0.001$ concentration. The longest values of T_2 are found in the vicinity of the gap minima (at $B_{0z} = B_{\min}$) for the smallest crystals (see Figs 2b, 3 and Methods), with values ranging from 5.2 μs to 8.4 μs, whereas the values are substantially shorter away from the minima. In fact, the T_2 values exhibit sharp divergences right at $B_{\min}$ (Fig. 3). The key to understanding this behaviour is the quadratic field dependence of the EPR transition frequencies (f) close to the gap minima (see Methods)

$$f = \Delta + \frac{\gamma_z^2}{2\Delta}(B_{0z} - B_{\min})^2 \tag{2}$$

such that the derivative $\mathrm{d}f/\mathrm{d}B_{0z} \propto (B_{0z} - B_{\min}) \to 0$ as $B_{0z} \to B_{\min}$; here, γ_z is the z component of the gyromagnetic tensor. Although not explicitly included in equation (1), nearly all sources of dipolar decoherence (due, for example, to dynamics associated with the nuclear bath and collective electron spin excitations, or magnons) can be approximated as a time-dependent magnetic noise, $\delta B_0(t)$, acting on the central spin qubit (the spin being measured) via the Zeeman interaction. In other words, processes that flip nearby spins cause variations in the local field, δB_0, at the position of the central spin, thereby altering its frequency/phase. Many of these processes involve indirect pairwise spin flip-flops (spin diffusion) that are extremely hard to mitigate, and persist to very low temperatures. The extreme axial anisotropy of HoW_{10} results in an insensitivity to the perpendicular applied field component, $B_{0\perp}$ (see Methods). Meanwhile, sensitivity to $\delta B_{0z}(t)$ vanishes (to first order) as $B_{0z} \to B_{\min}$ and $\mathrm{d}f/\mathrm{d}B_{0z} \to 0$, resulting in a vanishing contribution to the dipolar decoherence. This is the concept behind 'atomic clock transitions'. Named after the principle which gives atomic clocks their exceptional phase

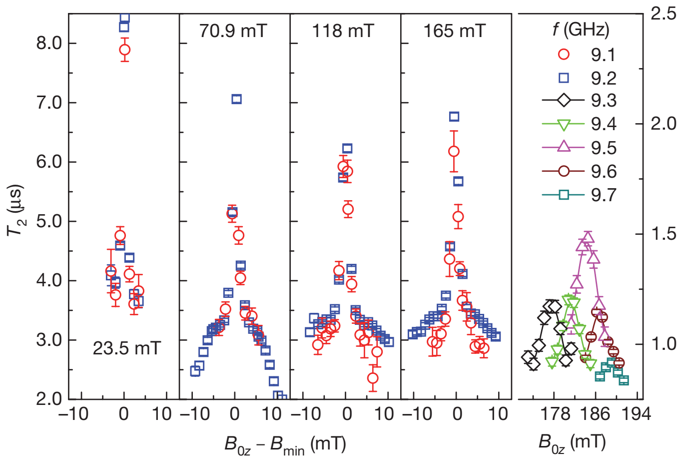


Figure 3 | T_2 divergence at the CTs. Field-swept T_2 measurements recorded at 5.0 K for a small $x = 0.001$ crystal at $\theta = 22°$ and various frequencies indicated in the key in the rightmost panel. The first four panels illustrate the divergences in T_2 at the CTs, referenced to the left-hand ordinate; the data are plotted in an expanded view as a function of $(B_{0z} - B_{min})$, with best-fit B_{min} values given in each panel. The rightmost panel, meanwhile, displays T_2 values well away from the CTs (see Fig. 1b), referenced to the right-hand ordinate. Error bars denote the standard error in T_2.

stability, these transitions are protected against environmental noise, resulting in dramatically enhanced quantum coherence[26,27]. Indeed, one expects the dephasing time, T_2, to scale as $(B_{0z} - B_{min})^{-n}$ ($n > 0$, see Extended Data Fig. 1)[28,29], thus explaining the observed divergences at the clock transitions (CTs). For comparison, T_2 measurements are displayed in the rightmost panel of Fig. 3 for several 'normal' EPR transitions, that is, $m_J = -4$ to $+4$ transitions away from the gap minima, where the frequency dependence approaches the linear regime and $df/dB_{0z} \to \gamma_z = 139.9\,\mathrm{GHz\,T^{-1}}$ (Fig. 1). Although T_2 is moderately peaked at the centres of these resonances, the sharp divergences seen at the CTs are clearly absent (see Methods for further discussion).

ESE-detected measurements for an $x = 0.01$ sample reveal divergences in T_2 at the CTs that are essentially identical to those seen in Fig. 3, with maximum values ranging from 4 μs to 8 μs (see Extended Data Figs 1 and 2). However, T_2 values associated with 'normal' EPR transitions well away from the CTs are much shorter (~100 ns, not shown). Because the collection of ESE spectra requires the detection of an echo, the observation of these 'normal' EPR transitions is challenging for $x \geq 0.01$. These findings are consistent with the idea that dipolar 'noise' increases with increasing Ho concentration, resulting in shorter T_2s for the 'normal' EPR transitions, yet there is an apparent insensitivity of T_2 to the Ho concentration right at the CTs.

Figure 4 displays 5 K ESE-detected spectra for a concentrated $x = 0.1$ sample that are in stark contrast to those in Fig. 2: narrow resonances are observed at the CTs that do not shift at all with frequency, that is, the data do not follow the simulations even though CW measurements indicate no measurable variation in the spin Hamiltonian parameters with Ho concentration[24]. The total suppression of 'normal' EPR transitions is attributed to a further reduction of T_2 on increasing the Ho concentration, to the extent that an echo can no longer be detected. Nevertheless, the T_2 values at the CTs remain long (~0.7 μs), resulting in the narrow ESE-detected resonances. Indeed, because the echo intensity is T_2-weighted, the resonance lineshape is a direct manifestation of the field dependence of T_2 at B_{min}. Analysis of CW EPR spectra suggests that the main contribution to the linewidth is a Gaussian distribution in the B_4^4 parameter ($\sigma_{B44} = 0.63$ MHz). This causes significant vertical broadening of the tunnelling gap, Δ, and EPR transition frequencies, as illustrated in Fig. 1b, which includes contours at the $\pm\sigma_\Delta$ and $\pm 2\sigma_\Delta$ levels ($\sigma_\Delta = 123$ MHz is the standard deviation in Δ; see Methods). These simulations indicate measurable intensity at

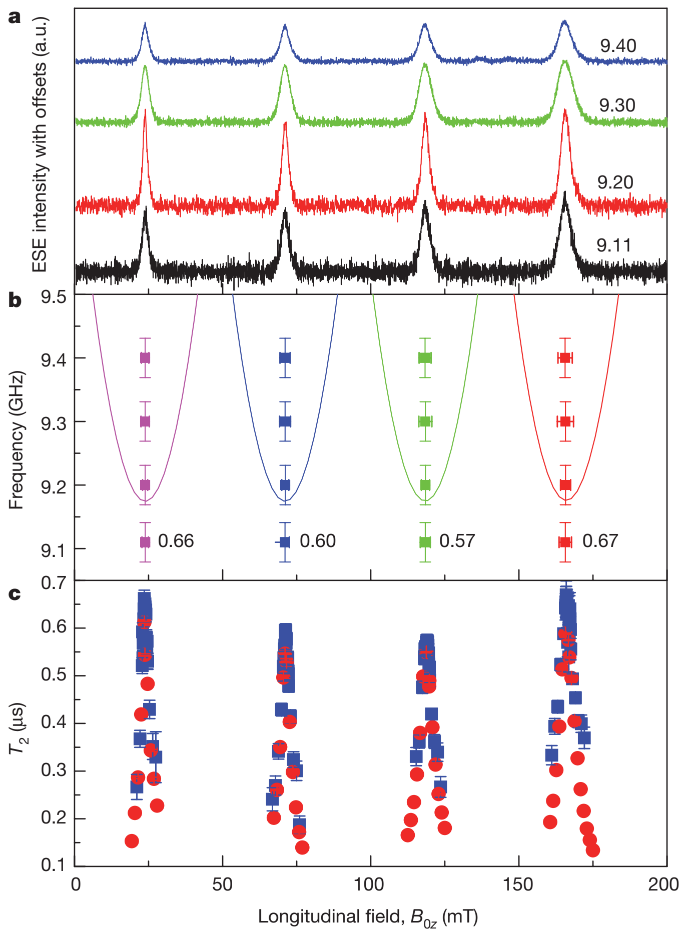


Figure 4 | ESE-detected spectra for a concentrated sample. a, Variable frequency measurements at 5.0 K for an $x = 0.10$ crystal, with $\theta = 20°$; the frequencies are indicated in GHz above each trace. The ESE resonances are attributed to CTs. a.u., arbitrary units. **b**, Frequency versus field plot of the CTs in **a**. Optimum T_2 values (in μs) are indicated next to the 9.11 GHz data. Meanwhile, the curves correspond to predictions based on the CW EPR parameterization[24]. **c**, Field-swept T_2 measurements recorded at 5.0 K for a separate $x = 0.10$ crystal at $\theta = 25°$ and frequencies of 9.12 GHz (blue squares) and 9.20 GHz (red circles). Vertical error bars in **b** denote pulse excitation bandwidths ($\pm 1/2\tau_{\pi/2}$, where $\tau_{\pi/2}$ is the duration of the π/2 pulse), while horizontal error bars represent standard deviations (±s.d.) deduced from Gaussian fits to the resonances in **a**. Error bars in **c** denote the standard error in T_2.

the CTs up to at least 9.4 GHz. However, the B_4^4 distribution does not shift the CTs appreciably to lower or higher fields, that is, all molecules in the distribution have their CTs at essentially the same B_{min} values. This explains the observation of narrow CT peaks spanning a wide frequency range in the $x = 0.1$ sample (Fig. 4); similar behaviour is also discernible at other concentrations (see Extended Data Fig. 3).

After magnetic 'noise', other sources of decoherence remain. First, the CTs do not protect against direct flip-flop processes that involve the central spin qubit[28,29]. These energy-conserving events involve coupling to other spins via the off-diagonal component of the dipolar interaction ($\hat{J}_1^+\hat{J}_2^- + \hat{J}_1^-\hat{J}_2^+$). The inhomogeneous broadening will provide some protection against this source of dephasing, because it requires the central spin to be resonant with other spins. Nevertheless, direct flip-flops probably explain the shorter T_2s at the CTs in the $x = 0.1$ sample. However, unlike the aforementioned indirect spin diffusion processes, direct flip-flops can be controlled at the stage of device design through the tuning/detuning of individual CT frequencies. Second, coupling of the Ho spin to lattice dynamics (phonons) via the CF is also likely to provide significant decoherence pathways, particularly as the temperature is raised[9,16]. Indeed, a significant

temperature dependence of T_2 is found at the CTs (a decrease by a factor of more than 2 upon heating the sample to 7 K), suggesting that T_2 may become limited by the spin–lattice relaxation time, T_1 ($\sim$20 μs at 5 K). This is something that will be the subject of future investigations.

The critical result from this study is the demonstration that CTs can be employed as a means of enhancing the coherence of molecular spin qubits in concentrated samples. Therefore, instead of attempting to suppress magnetic noise, which can be impractical at the stage of device design, we have shown here that one can fortify the molecular spin qubit itself against this noise through the use of CTs. In terms of design criteria, the molecule of choice should possess a large tunnelling gap within the ground magnetic doublet that matches the working frequency of the EPR cavity. The key to this strategy is the chemical design of molecular structures with appropriate CF states. In rare-earth complexes with integer spin, this goal translates into matching the m_J components of the ground doublet with the rotational order (q) of the main symmetry axis of the molecule (see Methods). Although this is not trivial to achieve, the case of HoW_{10} is not an isolated example. For example, within rigid polyoxometalate chemistry, the terbium derivative of the $[LnP_5W_{30}O_{110}]^{12-}$ series (where Ln indicates lanthanide) with pentagonal structure (approximate C_{5v} symmetry) has been characterized as having an $m_J = \pm 5$ ground state with an even larger tunnelling gap of $\sim$21 GHz that may be suitable for pulsed Q-band EPR[30]. Tunability across this range (10–100 GHz) is desirable and practical for quantum information applications, given that it matches the clock rate of state-of-the-art microprocessors. Moreover, operation at these CTs requires application of only very moderate magnetic fields (<0.2 T in the present example). Of course, this strategy can and should be combined with other ideas that are already being applied with success, such as using rigid lattices with a low abundance of nuclear spins[16]. Nevertheless, it is remarkable that working with CTs offers the unique advantage of allowing long coherence times with high concentrations of molecular spin qubits. In fact, for other molecular spin qubit candidates, T_2 values of the order of tens of microseconds were only observable in deuterated and highly diluted samples of Cr_7Ni molecular wheels[10] and $Cu(mnt)_2$ (mnt = maleonitriledithiolate) complexes[16].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 September 2015; accepted 5 January 2016.

1. Schlosshauer, M. A. *Decoherence and the Quantum-To-Classical Transition* (Springer, 2008).
2. Stamp, P. C. E. Quantum information: stopping the rot. *Nature* **453,** 167–168 (2008).
3. Devoret, M. H. & Schoelkopf, R. J. Superconducting circuits for quantum information: an outlook. *Science* **339,** 1169–1174 (2013).
4. Duan, L. Quantum physics: a strong hybrid couple. *Nature* **508,** 195–196 (2014).
5. Weitenberg, C. *et al.* Single-spin addressing in an atomic Mott insulator. *Nature* **471,** 319–324 (2011).
6. Pla, J. J. *et al.* A single-atom electron spin qubit in silicon. *Nature* **489,** 541–545 (2012).
7. Taminiau, T. H., Cramer, J., van der Sar, T., Dobrovitski, V. V. & Hanson, R. Universal control and error correction in multi-qubit spin registers in diamond. *Nature Nanotechnol.* **9,** 171–176 (2014).
8. Ardavan, A. *et al.* Will spin-relaxation times in molecular magnets permit quantum information processing? *Phys. Rev. Lett.* **98,** 057201 (2007).
9. Takahashi, S. *et al.* Decoherence in crystals of quantum molecular magnets. *Nature* **476,** 76–79 (2011).
10. Kaminski, D. *et al.* Quantum spin coherence in halogen-modified Cr_7Ni molecular nanomagnets. *Phys. Rev. B* **90,** 184419 (2014).
11. Wedge, C. J. *et al.* Chemical engineering of molecular qubits. *Phys. Rev. Lett.* **108,** 107204 (2012).
12. Leuenberger, M. N. & Loss, D. Quantum computing in molecular magnets. *Nature* **410,** 789–793 (2001).
13. Warner, M. *et al.* Potential for spin-based information processing in a thin-film molecular semiconductor. *Nature* **503,** 504–508 (2013).
14. Graham, M. *et al.* Influence of electronic spin and spin-orbit coupling on decoherence in mononuclear transition metal complexes. *J. Am. Chem. Soc.* **136,** 7623–7626 (2014).
15. Thiele, S. *et al.* Electrically driven nuclear spin resonance in single-molecule magnets. *Science* **344,** 1135–1138 (2014).
16. Bader, K. *et al.* Room temperature quantum coherence in a potential molecular qubit. *Nature Commun.* **5,** 5304 (2014).
17. Clemente-Juan, J. M. & Coronado, E. Magnetic clusters from polyoxometalate complexes. *Coord. Chem. Rev.* **193–195,** 361–394 (1999).
18. Clemente-Juan, J. M., Coronado, E. & Gaita-Ariño, A. Magnetic polyoxometalates: from molecular magnetism to molecular spintronics and quantum computing. *Chem. Soc. Rev.* **41,** 7464–7478 (2012).
19. Lehmann, L., Gaita-Ariño, A., Coronado, E. & Loss, D. Spin qubits with electrically gated polyoxometalate molecules. *Nature Nanotechnol.* **2,** 312–317 (2007).
20. van Hoogdalem, K., Stepanenko, D. & Loss, D. in *Molecular Magnets: Physics and Applications* (eds Bartolomé, J. *et al.*) 275–296 (Springer, 2014).
21. AlDamen, M. A., Clemente-Juan, J. M., Coronado, E., Martí-Gastaldo, C. & Gaita-Ariño, A. Mononuclear lanthanide single-molecule magnets based on polyoxometalates. *J. Am. Chem. Soc.* **130,** 8874–8875 (2008).
22. Martínez-Pérez, M. J. *et al.* Gd-based single-ion magnets with tunable magnetic anisotropy: molecular design of spin qubits. *Phys. Rev. Lett.* **108,** 247213 (2012).
23. AlDamen, M. A. *et al.* Mononuclear lanthanide single molecule magnets based on the polyoxometalates $[Ln(W_5O_{18})_2]^{9-}$ and $[Ln(\beta\text{-}SiW_{11}O_{39})_2]^{13-}$ (Ln^{III} = Tb, Dy, Ho, Er, Tm, and Yb). *Inorg. Chem.* **48,** 3467–3479 (2009).
24. Ghosh, S. *et al.* Multi-frequency EPR studies of a mononuclear holmium single-molecule magnet based on the polyoxometalate $[Ho^{III}(W_5O_{18})_2]^{9-}$. *Dalton Trans.* **41,** 13697–13704 (2012).
25. Schweiger, A. & Jeschke, G. *Principles of Pulse Electron Paramagnetic Resonance* (Oxford Univ. Press, 2001).
26. Bollinger, J. J. *et al.* Laser-cooled-atomic frequency standard. *Phys. Rev. Lett.* **54,** 1000–1003 (1985).
27. Vion, D. *et al.* Manipulating the quantum state of an electrical circuit. *Science* **296,** 886–889 (2002).
28. Wolfowicz, G. *et al.* Atomic clock transitions in silicon-based spin qubits. *Nature Nanotechnol.* **8,** 561–564 (2013).
29. Balian, S. J., Wolfowicz, G., Morton, J. J. L. & Monteiro, T. S. Quantum bath-driven decoherence of mixed spin systems. *Phys. Rev. B* **89,** 045403 (2014).
30. Cardona-Serra, S. *et al.* Lanthanoid single-ion magnets based on polyoxometalates with a 5-fold symmetry: the series $[LnP_5W_{30}O_{110}]^{12-}$ (Ln^{3+} = Tb, Dy, Ho, Er, Tm, and Yb). *J. Am. Chem. Soc.* **134,** 14982–14990 (2012).

Acknowledgements We thank L. Song and J. van Tol for technical assistance with the X-band EPR spectrometer. This work was supported by the NSF (grant DMR-1309463) and the US AFOSR (AOARD contract 134031 FA2386-13-1-4029). Work performed at the NHMFL was supported by the NSF (DMR-1157490) and by the State of Florida. Work performed at the Instituto de Ciencia Molecular was supported by the European Research Council (grants SPINMOL and DECRESIM), by the Spanish MINECO (projects MAT-2014-56143-R, CTQ2014-52758-P and Excellence Unit Maria de Maeztu MDM-2015-0538) and by the Generalidad Valenciana (Prometeo and ISIC-Nano Programs of Excellence). A.G.-A. thanks the Spanish MINECO for a Ramón y Cajal Fellowship.

Author Contributions S.H., E.C. and A.G.-A. conceived the research and wrote the paper. Y.D. prepared the samples. S.H., M.S. and D.K. designed the experiments, while M.S. and D.K. performed the measurements. S.H., M.S. and D.K. analysed the results.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.H. (shill@magnet.fsu.edu) or E.C. (eugenio.coronado@uv.es).

METHODS

Experimental details. Pulsed EPR measurements were performed on a commercial Bruker E680 X-band spectrometer equipped with a cylindrical TE_{011} dielectric resonator (model ER 4118 X-MD5) with a centre frequency $f_0 = 9.75$ GHz. Single-crystals of $Na_9[Ho_xY_{(1-x)}(W_5O_{18})_2]\cdot nH_2O$ ($x = 0.001$ to 0.25) were prepared according to the method described in ref. 23. Samples were re-crystallized before study, then transferred to the spectrometer directly from the mother liquor and cooled rapidly in order to prevent loss of crystallinity due to evaporation of lattice solvent. The sample temperature was controlled using an Oxford Instruments CF935 helium flow cryostat and ITC503 temperature controller. A strong temperature dependence of T_2 at the CTs required operation of the cryostat at a temperature of 5.0 K in order to ensure good thermal stability and sample-to-sample reproducibility.

For each series of measurements, a single crystal was mounted on a 4 mm diameter quartz rod and positioned at the centre of the cylindrical resonator for perpendicular mode excitation. The tendency for samples to rapidly lose solvent, and the low symmetry $P\bar{1}$ space group of the HoW_{10} compound, made it impossible to index and align crystals before mounting. However, the Bruker E680 and ER 4118X-MD5 dielectric resonator combination allows for *in situ* sample rotation about a single axis. Each crystal was therefore aligned as best as possible on the basis of angle-dependent CW EPR measurements performed at 9.75 GHz and 5.0 K. The remaining misalignment, θ, between $\boldsymbol{B}_0$ and z was determined by scaling the applied field to match the simulations in Fig. 1 (see below). A $\theta < 30°$ criterion was then applied; crystals not meeting this condition were discarded and a new sample selected for study.

When overcoupled for ESE measurements, the bandwidth of the resonator is given by $\Delta f = f_0/Q \approx 250$ MHz, where the loaded quality factor $Q \approx 40$. This is sufficient to allow variable-frequency measurements with reasonable microwave B_1 fields down to a lower limit of ~9.1 GHz. The B_1 fields were independently measured under the same conditions via the Rabi oscillation frequency (Ω_R) of a spin-1/2 EPR standard (the organic radical bisdiphenylene-2-phenylallyl dissolved in polystyrene); B_1 values varied from ~4 G at 9.1 GHz, to 9 G at 9.75 GHz ($\Omega_R = 11$–25 MHz for $s = 1/2$). A two-pulse sequence ($T/2$–τ–T–τ–echo, where T characterizes the pulse durations and τ the delay time between pulses) was employed for all ESE measurements reported in this work. The values of T, τ and the source power were optimized at each frequency, with the assumption that the optimum conditions correspond approximately to the Hahn-echo sequence, $\pi/2$–τ–π–τ–echo, where π refers to the tipping angle. For T_2 measurements, τ was varied and the resultant echo amplitude then fitted to a single exponential decay.

Pulse sequences. Because the ESE measurements were performed well below the centre frequency of the cavity, and owing to the lack of a priori knowledge of the matrix elements associated with the observed transitions, pulse sequences were adjusted at each frequency by one of two methods: (1) the $\pi/2$ pulse length ($T/2$) and source attenuation were adjusted to maximize the echo intensity relative to the spectrometer noise for the ESE-detected spectra in Figs 2a and 4a, thereby explaining the variability of the vertical error bars denoting excitation bandwidth (defined as $2/T$, or $1/\tau_{\pi/2}$, where $\tau_{\pi/2}$ is the duration of the $\pi/2$ pulse in the Hahn-echo sequence); and (2) Rabi oscillation measurements were used to determine the optimum $\pi/2$ pulse length for the detailed T_2 measurements displayed in Figs 3 and 4c, and Extended Data Figs 1 and 2 (the Rabi pulse sequence was optimized via method (1)). On this basis, a Rabi frequency, $\Omega_R = 98\ \mathrm{Mrad\,s^{-1}}$ (15.6 MHz), was determined for 0 dB attenuation at the CTs, resulting in a minimum $\pi/2$ pulse length of 16 ns for the employed spectrometer. This corresponds to an optimum dephasing factor $Q_\varphi = 820$, defined here as $Q_\varphi = \Omega_R T_2$, a figure of merit for qubit operation. We note, however, that this does not preclude shorter pulses using a more powerful microwave source, suggesting the possibility of Q_φ values up to 1.5×10^6 using the modified definition in ref. 9. Interestingly, this value is identical to the one reported in ref. 9 for an Fe_8 nanomagnet, in spite of the vastly different frequencies employed in the two measurements, primarily because of the much longer coherence in the HoW_{10} system. Based on knowledge of the spectrometer used for the Fe_8 study, we estimate a $Q_\varphi = \Omega_R T_2$ of just 50 for Fe_8; of course, the same arguments concerning limited source power apply in that case. The HoW_{10} Q_φ value compares favourably with other candidate molecular spin qubits using both definitions—for example, the optimum Q_φ ($= \Omega_R T_2$) varies from ~2,000 for the Cr_7Ni wheel (ref. 10), up to ~10,000 obtained recently for a Cu^{II} coordination complex[16]. However, one should bear in mind that extreme dilution/deuteration was employed in these cases.

The spin Hamiltonian. The energy spectrum associated with the Hund's rule spin–orbit coupled ground state of the Ho^{3+} ion, with $L = 6$, $S = 2$ and $J = |\boldsymbol{L} + \boldsymbol{S}| = 8$, can be described by the effective Hamiltonian (equation (1) in the main text, reproduced here for convenience)

$$\hat{H} = \sum_{k=2,4,6} \sum_{q=0}^{k} B_k^q \hat{O}_k^q + \hat{J} \cdot A \cdot \hat{I} + \mu_B \boldsymbol{B}_0 \cdot g_L \cdot \hat{J} - \mu_N g_N \boldsymbol{B}_0 \cdot \hat{I}$$

The double summation describes the CF interaction in terms of extended Stevens operators $\hat{O}_k^q$ ($k = 2, 4, 6$, and $|q| \le k$), with associated coefficients B_k^q (refs 31, 32), and with $\hat{O}_k^q$ expressed in terms of the total electronic angular momentum operators $\hat{J}$ and $\hat{J}_i$ ($i = x, y, z$). Using this convention, the axial ($q = 0$) coefficients determined from magnetic and continuous-wave (CW) EPR measurements are[23,24]: $B_2^0 = 0.601\ \mathrm{cm^{-1}}$, $B_4^0 = 6.96 \times 10^{-3}\ \mathrm{cm^{-1}}$, and $B_6^0 = -5.10 \times 10^{-5}\ \mathrm{cm^{-1}}$. This parameterization results in the $m_J = \pm 4$ CF states lying lowest in energy (Fig. 1), separated from the $m_J = \pm 5$ excited states by ~20 $\mathrm{cm^{-1}}$ (ref. 23). The second term in equation (1) describes the hyperfine coupling between the Ho^{3+} electron and $I = 7/2$ nuclear spin, resulting in the observation of eight ($2I + 1$) well-resolved electro-nuclear transitions via high-field CW EPR measurements; here, $\hat{I}$ denotes the total nuclear angular momentum operator, and A the hyperfine coupling tensor, for which the parallel component, $A_\parallel = 830 \pm 10$ MHz, has been determined from the high-field CW EPR spectrum[24]. The final two terms in equation (1) respectively parameterize the electron and nuclear Zeeman interactions with the local magnetic induction, $\boldsymbol{B}_0$, in terms of a Landé g-tensor (g_L) and isotropic nuclear g-factor (g_N); μ_B and μ_N represent the Bohr (electron) and nuclear magneton, respectively. The parallel component of the Landé g-tensor, $g_z = 1.25(1)$, has been determined from CW EPR studies[24].

In addition to the axial ($q = 0$) CF parameters, CW EPR measurements at X-band frequencies can only be accounted for by including a sizeable tetragonal $B_4^4\hat{O}_4^4$ ($\hat{O}_4^4 = \frac{1}{2}(\hat{J}_+^4 + \hat{J}_-^4)$) interaction, with $B_4^4 = 3.14 \times 10^{-3}\ \mathrm{cm^{-1}}$ (see Fig. 1a and ref. 24 for detailed explanation). It is this term (which is allowed because of a small distortion of the HoW_{10} molecule away from exact D_{4d} symmetry) that generates avoided level crossings between $m_J = \pm 4$ states, as seen in Fig. 1a. In principle, the sixth order tetragonal $B_6^4\hat{O}_6^4$ interaction is also symmetry allowed. However, $\hat{O}_6^4$ contains the commutator $[\hat{J}_z^2, (\hat{J}_+^4 + \hat{J}_-^4)]$ and is, thus, indistinguishable from $\hat{O}_4^4$ within the truncated $m_J = \pm 4$ ground doublet. Therefore, we employ only the $B_4^4\hat{O}_4^4$ term to capture the effects of the distortion away from exact D_{4d} symmetry. The key point is that $\hat{O}_4^4$ connects the $m_J = \pm 4$ states in second-order, resulting in unusually large (~9 GHz) quantum tunnelling gaps. For $\boldsymbol{B}_0 \parallel z$, the frequencies of the resultant weakly allowed EPR transitions between these states then follow a field-dependence of the form (see Fig. 1b)

$$f = \sqrt{\Delta^2 + \gamma_z^2 (B_{0z} - B_{\min})^2} \approx \Delta + \frac{\gamma_z^2}{2\Delta}(B_{0z} - B_{\min})^2 \qquad (3)$$

where the approximate quadratic expression applies for fields close to the gap minima, $B_{\min}$. Indeed, because $\hat{O}_4^4$ represents the only off-diagonal CF interaction in equation (1), an almost exact mapping of the first expression of equation (3) onto curves generated via exact diagonalization of equation (1) is possible, yielding the following parameters: $\Delta = 9.18$ GHz, $\gamma_z = 139.9\ \mathrm{GHz\,T^{-1}}$ ($= 1.25 \times 8 \times \mu_B/h$, that is, $g_z = 1.25$), and $B_{\min} = 23.6$, 70.9, 118.1 and 165.4 mT. This analysis assumes $\boldsymbol{B}_0 \parallel z$, while the experiments are typically performed with a small field misalignment ($\theta \ne 0$), as noted above. However, due to the extreme uniaxial symmetry of the HoW_{10} molecule, the perpendicular component of the effective gyromagnetic tensor associated with the $m_J = \pm 4$ doublet, $\gamma_{\perp,\mathrm{eff}} < 0.1\ \mathrm{GHz\,T^{-1}}$ ($g_{\perp,\mathrm{eff}} < 0.01$), resulting in a virtual insensitivity to the perpendicular component of the applied field ($B_{0\perp}$) over the range explored in this investigation; for comparison, note that $\gamma_{\mathrm{proton}} \approx 0.04\ \mathrm{GHz\,T^{-1}}$. For this reason, one can approximate the electronic Zeeman term in equation (1) using a scalar interaction of the form, $g_z \mu_B B_{0z} \hat{J}_z$ (where $B_{0z} = B_0 \cos\theta$). Equation (3) then applies quite generally at the gap minima, provided the applied field is rescaled to account for any misalignment. Hence all EPR spectra are plotted as a function of the longitudinal applied field component, B_{0z}. Importantly, the derivative $\mathrm{d}f/\mathrm{d}B_{0z} \to 0$ (that is, $\gamma_{z,\mathrm{eff}} \to 0$) as $B_{0z} \to B_{\min}$, resulting in an almost complete insensitivity of the EPR transition frequencies at the gap minima to magnetic noise associated with the environment, thus giving rise to the strong T_2 divergences at the CTs. However, the small yet finite $\gamma_{\perp,\mathrm{eff}}$ ($< 0.1\ \mathrm{GHz\,T^{-1}}$) probably limits T_2 right at the CTs (within ± 0.5 G of $B_{\min}$) in these studies due to the unavoidable field misalignment. In fact, $\gamma_{\perp,\mathrm{eff}} \to 0$ as $B_0 \sin\theta \to 0$, which may explain the longer T_2 values observed at the lowest field CTs in Fig. 3, and also suggests that longer T_2s may be achievable in precisely aligned samples.

T_2 scaling. The data displayed in Fig. 3 were obtained for a small crystal of the most dilute sample ($x = 0.001$). It is the high quality of this crystal that results in the sharp T_2 peaks at all four CTs (all four $B_{\min}$ locations). However, it gives weak ESE signals, making it challenging to perform a detailed analysis of the scaling of T_2 with B_{0z}. Careful T_2 measurements were therefore repeated for larger samples. Unfortunately, the larger crystals are susceptible to twinning that manifests as a broadening of spectral peaks and T_2 divergences, with the effect being more pronounced at the higher field CTs (see Fig. 2a). However, the first CT at $B_{\min} = 23.6$ mT often remains sharp (see below for explanation). Extended Data Fig. 1 displays T_2 measurements for the $x = 0.001$ and 0.01 concentrations, plotted

against $(B_{0z} - B_{min})$ on both logarithmic (main panels) and linear (insets) scales. Similar to the data in Fig. 3, the T_2 peaks exhibit broad tails, with an apparent kink at $|B_{0z} - B_{min}| \approx 2$ mT for the more dilute sample. However, when plotted on a log-log scale, the data follow a power law (to within the experimental uncertainty) spanning an order of magnitude in $(B_{0z} - B_{min})$ for $x = 0.001$, and almost two orders of magnitude for $x = 0.01$, particularly on the high-field sides of the T_2 peaks. This apparent monotonic behaviour of the form $T_2 \propto (B_{0z} - B_{min})^{-n}$ supports our assertion that the decoherence is dominated by dipolar field fluctuations that vanish as $df/dB_{0z} \to 0$. However, the exponent, n, is both sample-dependent ($n = 0.33$ and 0.46, respectively, for $x = 0.001$ and 0.01), and different from previous predictions[28,29]: $n = 1$ for indirect flip-flop processes (spin diffusion), and $n = 2$ for instantaneous diffusion[25]. We believe that sample inhomogeneity is responsible for these differences in HoW_{10}, thus masking the intrinsic T_2 dependence on B_{0z}, causing obvious sample-to-sample variability. It is nevertheless interesting that a power-law scaling still holds, as opposed, for example, to Gaussian behaviour. This clearly merits further theoretical investigation.

Reduced ESE intensity and faster T_2 decay curves are part of the reason for the increased error bars and apparent broad tails seen in Fig. 3 and Extended Data Fig. 1. In addition, ESE-envelope-modulation (ESEEM)[25] is detectable in the decay curves recorded in these tails (not shown). However, only one to two heavily damped periods of oscillation can be seen, thus adding to the error in T_2 (not to mention a potential systematic error that is not taken into account in our analysis). It is these combined factors that likely explain the apparent kink in some of the data at $|B_{0z} - B_{min}| \approx 2$ mT, as well as the weak variation in T_2 across the 'normal' transitions seen in the right-hand panel of Fig. 3. Interestingly, enough ESEEM periods can be detected to confirm that it is due to coupling to protons in the sample. Importantly, the ESEEM vanishes at the CTs, providing further strong evidence that the Ho^{3+} spin becomes decoupled from the surrounding dipolar spin bath as both $(B_{0z} - B_{min})$ and $df/dB_{0z} \to 0$.

Spectral broadening. The EPR spectra of HoW_{10} are inhomogeneously broadened[24], with the two main contributions originating from (i) crystal twinning and (ii) strain in the off-diagonal B_4^4 CF parameter.

(i) Crystals of HoW_{10} form as long thin needles that tend to aggregate into aligned bundles. Separating single crystals from these bundles can be challenging, particularly given that removal of the samples from their mother liquor for periods of more than a few minutes leads to sample degradation. Even after separation, our measurements suggest varying degrees of mosaic spread, particularly for the larger crystals. Indeed, simulations of high-field CW EPR spectra (where the effects of the mosaicity are more pronounced than at X-band) employed a Gaussian orientational distribution with a full-width at half-maximum (FWHM) of 1°, albeit for a small crystal[24]; the distribution is considerably broader for many of the samples employed for ESE measurements. Within the context of equation (2), this mode of disorder produces a spread in γ_z and the B_{min} values, resulting in horizontal smearing of the energy levels in Fig. 1, as opposed to a vertical smearing produced by a distribution in B_4^4 (see below). The horizontal smearing becomes more pronounced at higher fields, akin to g-strain. Consequently, the EPR spectra often become broader with increasing field, as is clearly evident in Fig. 2, and less so in Fig. 4.

Although subtle, the effects of sample mosaicity are most pronounced at the CTs. The horizontal spread in the CTs results in a smearing of the divergence in T_2. In general, the strongest/narrowest divergences were obtained for the smallest crystals, which have the smallest mosaic spread. It is for this reason that the data for the most dilute samples in Figs 2 and 3 were obtained for two different crystals: the large crystal employed in Fig. 2a did not produce particularly strong T_2 divergences, with maximum values reaching only ~2 μs. Meanwhile, a smaller crystal was employed in Fig. 3: this sample gave very good echoes right at the CTs, in spite of its reduced spin count; however, its ESE spectra vanish into the noise upon moving appreciably away from the CTs. These trends can be attributed both to a T_2 weighting effect, which amplifies the otherwise weak ESE signals at the CTs for the more ordered (longer T_2) sample, and to the narrower mosaic distribution that further enhances echoes at the CTs. Multiple small samples were studied, and optimum T_2 values at the CTs in the 6–8 μs range were found in nearly all cases for the $x = 0.001$ and 0.01 samples (see Fig. 3 and Extended Data Figs 1 and 2).

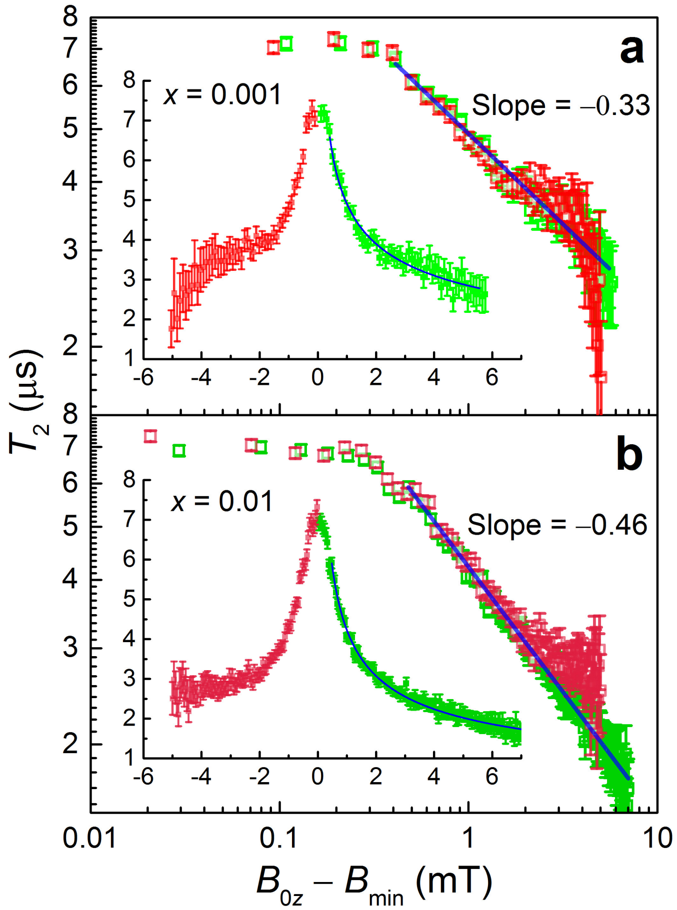
(ii) Other sources of inhomogeneous broadening include: strains in the spin Hamiltonian parameters (B_k^q, A and g_L), caused by microscopic disorder, and inhomogeneities in $\boldsymbol{B}_0$ due to electron and nuclear dipolar fields. The latter may be ruled out as a major source of broadening at X-band (and 5 K) due to weak sample magnetization and the lack of any systematic dependence of the EPR linewidth on Ho concentration. Meanwhile, the only effect of the diagonal ($q = 0$) CF terms in equation (1) is to ensure an isolated $m_J = \pm 4$ doublet ground state with $\gamma_z = g_J J \mu_B / h = 139.9\,\mathrm{GHz\,T^{-1}}$ ($J = 8$ and $g_J = 1.25$). Other than that, the low energy spectrum exhibits little or no dependence on B_2^0, B_4^0 and B_6^0 (ref. 24), and should thus be insensitive to strains in these parameters. For related reasons, and because of the contracted nature of the $4f$ shell and strong spin-orbit coupling, the g_L and A tensors are relatively immune to local strains in the crystal structure (although the effective interactions will of course be sensitive to sample alignment due to the strong axial character of the CF). This leaves B_4^4 which, indeed, has a profound influence on the X-band EPR spectrum, as clearly seen in Fig. 1a and discussed in detail in ref. 24: B_4^4 directly sets the scale of the tunnelling gap, Δ, which is responsible for the CTs.

The finite B_4^4 parameter arises because of a small deviation of the coordination environment around the Ho ion from exact D_{4d} symmetry[24]. The superposition of disorder onto this weakly distorted structure can then give rise to a relatively strong modulation of the local B_4^4 parameter and, hence, to a broad distribution for the ensemble. Working under this assumption, we re-simulated CW X-band spectra obtained for an $x = 0.1$ sample at a frequency of 9.64 GHz (figure 8 of ref. 24), assuming that the main source of broadening is a Gaussian distribution in B_4^4. The best simulation is obtained with a FWHM of $5.0 \times 10^{-5}\,\mathrm{cm^{-1}}$, that is, ~1.6% of B_4^4 (or a standard deviation, $\sigma_{B44} = 2.1 \times 10^{-5}\,\mathrm{cm^{-1}}$). This, in turn, produces a vertical distribution in the corresponding tunnelling gap, Δ. Because $\hat{O}_4^4$ connects the $m_J = \pm 4$ states at the second order of perturbation, the resultant standard deviation of the gap distribution is given approximately by $\sigma_\Delta \approx 2\Delta\sigma_{B44}/B_4^4 = 4.1 \times 10^{-3}\,\mathrm{cm^{-1}} = 123$ MHz (FWHM of 290 MHz), where $\Delta = 0.306\,\mathrm{cm^{-1}} = 9.18$ GHz is the mean gap value (the factor of '2' emerges because of the quadratic dependence of Δ on B_4^4).

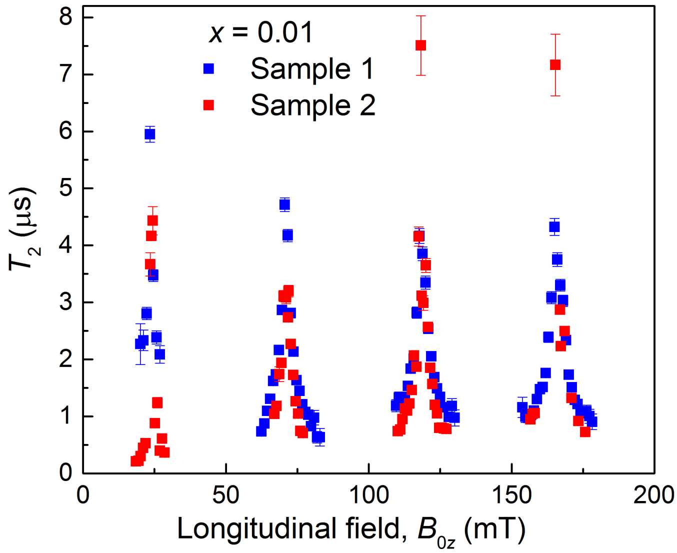
Figure 1b depicts the Gaussian broadening of the EPR transition frequencies as a 3D colour map, with contours shown at the $\pm\sigma_\Delta$ and $\pm 2\sigma_\Delta$ levels of the distribution. Because B_4^4 affects only Δ, this mode of disorder does not shift the magnetic fields (B_{min}) at which the CTs occur for the different molecules in the distribution. However, it does distribute them vertically over a relatively wide frequency range (approximately ± 0.25 GHz at the $2\sigma_\Delta$ level). This can explain the observation of ESE intensity exactly at the CTs over a wide frequency range for the concentrated ($x = 0.1$) sample seen in Fig. 4. Because the cavity employed for these investigations has a centre frequency at 9.75 GHz, its sensitivity improves upon increasing the frequency from 9.1 to 9.4 GHz. Meanwhile, the number of Ho^{3+} spins in the distribution decreases with increasing frequency. These two factors approximately offset, explaining the relatively constant ESE intensity and signal-to-noise ratio across the studied frequency range. The ESE intensity does peak at 9.2 GHz, above which it decays, although not as rapidly as one may expect purely on the basis of the gap distribution. This is due to the increasing B_1 field of the spectrometer, which enables excitation of more spins and hence the generation of stronger echoes at higher frequencies.

This same behaviour is observable at the other concentrations. For example, CTs are very clearly observable in between the 'normal' EPR transitions over a wide frequency range at $B_{0z} = 165$ mT for the $x = 0.01$ sample, as seen in Extended Data Fig. 3. Further evidence can also be found at some of the higher frequencies, where inspection of Fig. 1a reveals crossings between nuclear sub-levels ($\Delta m_I = \pm 1$) at fields exactly half way between the B_{min} values. If the applied field is not well aligned to the crystal z axis, these become avoided crossings (with <10 MHz gaps), giving rise to new CTs at these higher frequencies. This is a subtlety of the perpendicular field component, $B_{0\perp}$, which will be the subject of a future publication. The avoided nuclear sub-level crossings do not influence any of the conclusions concerning the CTs at the B_4^4 gap minima (Δ). Nevertheless, the higher frequency CTs are observable, particularly at low fields where the effects of disorder due to sample mosaicity are less pronounced, and the 'normal' ESE transitions are quenched due to very short T_2s (refs 2, 9). This is the explanation for the sharp double peaks seen for the $x = 0.001$ sample at ~50 mT between 9.4 and 9.7 GHz in Fig. 2a, as well as the sharp zero-field peaks and some of the fine structures seen between B_{min} values at higher fields and frequencies. On the basis of the 50 mT CTs, one can see that the vertical broadening spans less than 400 MHz in this sample, that is, less than ± 200 MHz from the peak of the distribution. In other words, σ_{B44} clearly varies from sample to sample, being smaller for the $x = 0.001$ concentration. This is the reason why intensity due to the low-frequency CTs (at B_{min}) is not discernible in between the broad 'normal' transitions in the most dilute sample in Fig. 2a.

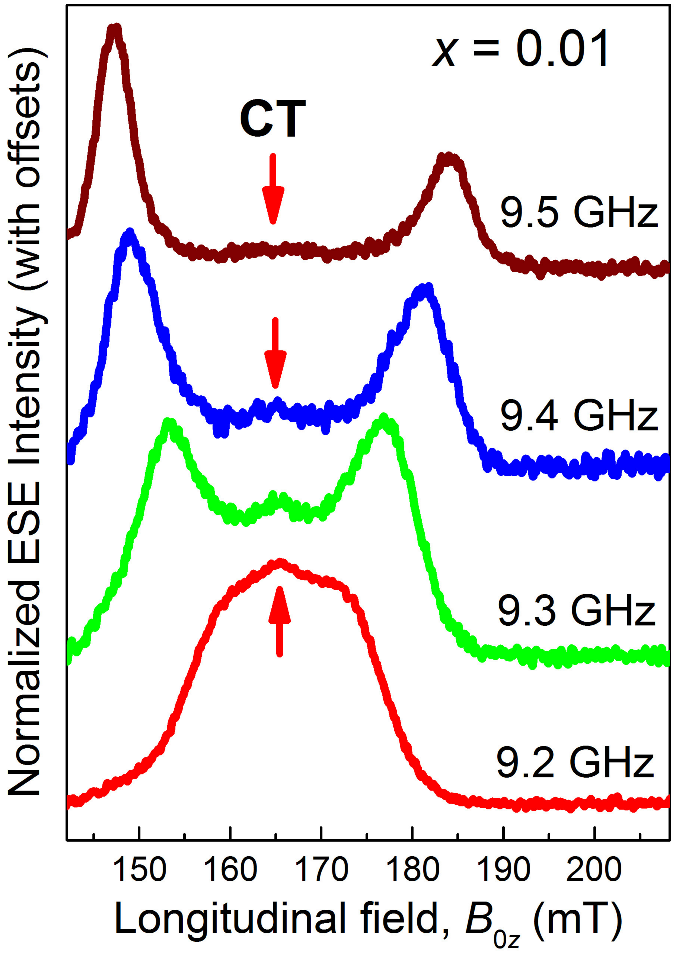
31. Rudowicz, C. & Chung, C. Y. The generalization of the extended Stevens operators to higher ranks and spins, and a systematic review of the tables of the tensor operators and their matrix elements. *J. Phys. Condens. Matter* **16,** 5825–5847 (2004).
32. Stoll, S. & Schweiger, A. EasySpin, a comprehensive software package for spectral simulation and analysis in EPR. *J. Magn. Reson.* **178,** 42–55 (2006).



Extended Data Figure 1 | T_2 scaling. a, **b**, Field-swept T_2 measurements for the $x = 0.001$ (**a**) and $x = 0.01$ (**b**) concentrations at 5 K; the data are plotted as a function of ($B_{0z} - B_{min}$) on both log–log (main panels) and linear (insets) scales. The blue lines are power-law fits to the positive ($B_{0z} - B_{min}$) data (green points), with the obtained exponents ('slope') given in the figures. Error bars, ±standard error in T_2.



Extended Data Figure 2 | T_2 divergence at the $x = 0.01$ concentration. Shown are field-swept T_2 measurements recorded at 5.0 K for two separate crystals at frequencies of 9.12 GHz (blue squares) and 9.20 GHz (red circles). Error bars, ±standard error in T_2.



Extended Data Figure 3 | ESE-detected spectra for the $x = 0.01$ concentration. Variable frequency measurements at 5.0 K, with $\theta = 30°$; the frequencies are indicated above each trace. Similar to spectra for the $x = 0.001$ sample, the broad 9.2 GHz CT peak splits into two upon moving away from the tunnelling gap minimum (see also Fig. 2). However, weak ESE intensity can still be detected at $B_{0z} = 165$ mT at all four frequencies. This is due to vertical broadening of the CT, caused by a Gaussian distribution in B_4^4.

doi:10.1038/nature17141

Three-dimensional control of the helical axis of a chiral nematic liquid crystal by light

Zhi-gang Zheng[1], Yannian Li[1], Hari Krishna Bisoyi[1], Ling Wang[1], Timothy J. Bunning[2] & Quan Li[1]

Chiral nematic liquid crystals—otherwise referred to as cholesteric liquid crystals (CLCs)—are self-organized helical superstructures that find practical application in, for example, thermography[1], reflective displays[2], tuneable colour filters[3,4] and mirrorless lasing[5,6]. Dynamic, remote and three-dimensional control over the helical axis of CLCs is desirable, but challenging[7,8]. For example, the orientation of the helical axis relative to the substrate can be changed from perpendicular to parallel by applying an alternating-current electric field[9], by changing the anchoring conditions of the substrate, or by altering the topography of the substrate's surface[10–16]; separately, in-plane rotation of the helical axis parallel to the substrate can be driven by a direct-current field[17–19]. Here we report three-dimensional manipulation of the helical axis of a CLC, together with inversion of its handedness, achieved solely with a light stimulus. We use this technique to carry out light-activated, wide-area, reversible two-dimensional beam steering—previously accomplished using complex integrated systems[20] and optical phased arrays[21]. During the three-dimensional manipulation by light, the helical axis undergoes, in sequence, a reversible transition from perpendicular to parallel, followed by in-plane rotation on the substrate surface. Such reversible manipulation depends on experimental parameters such as cell thickness, surface anchoring condition, and pitch length. Because there is no thermal relaxation, the system can be driven either forwards or backwards from any light-activated intermediate state. We also describe reversible photocontrol between a two-dimensional diffraction state, a one-dimensional diffraction state and a diffraction 'off' state in a bilayer cell.

According to Bragg's law, when CLCs are in a planar cell—and hence their helices are in a 'standing helix' (SH) orientation, perpendicular to the substrate—modulating the helical pitch length produces tuneable, selective reflection of circularly polarized light. In contrast, CLCs in a homeotropic cell—where the helical axes are in a 'lying helix' (LH) orientation, parallel to the substrate's surface, but randomly oriented—exhibit a fingerprint optical texture. Such an LH arrangement has allowed rotational manipulation of microscale objects on the surface of CLC films[22]. On the other hand, a 'uniform LH arrangement', in which the helical axes are oriented along a single direction, produces an optical texture of uniform periodic stripes perpendicular to the helical axis, and an in-plane, periodic modulation of the refractive index along the helical axis. Varying the pitch length of the uniform LH arrangement can modulate the diffraction angle, enabling non-mechanical beam steering and spectrum scanning along a one-dimensional line[9,23]. A wide in-plane rotation angle of the helical axis has been produced in a hybrid cell (with one substrate treated for vertical alignment, the other for homogeneous alignment) by light irradiation[16], but the helical axis could not be transformed from the LH to the SH state. Other work has used independent external stimuli to transform standing helices to lying helices, or to achieve in-plane rotation of uniform lying helices[9–19]. Here, we use light to induce both events sequentially—transformation of the SH to the uniform LH arrangement, followed by in-plane rotation—enabling three-dimensional control over the helical axis.

We used our synthesized, dithienylcyclopentene-based, axially chiral molecular switch—(*S*,*S*)-**D4** (Fig. 1a and Supplementary Information)—as the dopant (1.2 mol%) in the commercially available achiral nematic liquid crystal E7, to fabricate a self-organized, optically tuneable CLC. The helical twisting power (HTP) of (*S*,*S*)-**D4** in E7 has previously been determined in a wedge cell[24]; a texture transition of such a CLC confined in a homeotropic cell has also been observed[25], although no reversible SH-to-LH transition and in-plane rotation of the LH have previously been found. This axially chiral switch shows excellent fatigue resistance with superior thermal stability in both its ring-open and its ring-closed states[24]. Upon irradiation with ultraviolet light at 310 nm, the ring-open structure is transformed into the ring-closed isomer; the helical superstructure changes handedness, from the initial right-handed to a left-handed form; and the HTP is enhanced. The reverse process occurs upon irradiation with visible light at 550 nm.

We filled this photoresponsive CLC into a homogeneous planar cell (where the two substrates were aligned antiparallel to each other, and the cell gap, that is the gap between the top and bottom substrates, was $3.7 \pm 0.1\,\mu m$). We used a polarizing optical microscope in transmission mode to study the sample (Fig. 1b). Initially, it is in a bright state, indicating the expected Grandjean planar texture—standing helices. Upon irradiation with ultraviolet light for 5 seconds, the bright state transforms into a dark state, corresponding to the unwound nematic phase, resulting from the homogeneous alignment of liquid-crystal (LC) molecules (parallel to the polarization direction of incident light). After 10 seconds of irradiation, the bright state reappears (indicating the emergence of standing helices with opposite handedness), followed by the appearance of the periodic stripes that indicate a uniform LH arrangement, and accompanied by simultaneous in-plane rotation of the stripes and pitch contraction until the system reaches the photostationary state (PSS). In contrast with aforementioned work[7–17], here the uniform LH arrangement is formed only by light irradiation. This LH structure can be erased and driven reversibly with visible light irradiation, as follows: the stripes rotate in the opposite direction; the distance between two adjacent stripes increases; the helices align perpendicularly, producing a left-handed SH arrangement; this left-handed structure unwinds and reorganizes to produce the right-handed SH arrangement; and eventually the uniform LH texture of the right-handed CLC is regenerated (Fig. 1d). Thus, the direction of the helical axis of CLCs can be manipulated in three dimensions solely by light (Fig. 1c). Moreover, the CLC system in any stimulated intermediate state is stable, without showing thermal relaxation, because of the thermal stability of the chiral molecular switch in both of its isomeric states.

The light-induced uniform LH arrangement might be produced in two main ways: first, through development of a large oblique or a vertical alignment of the LC molecules (this would benefit LH formation by coupling with the chiral effects); and second, through sufficient surface anchoring to maintain the orientation of the stripes in a single direction

[1]Liquid Crystal Institute and Chemical Physics Interdisciplinary Program, Kent State University, Kent, Ohio 44242, USA. [2]Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson AFB, Ohio 45433, USA.

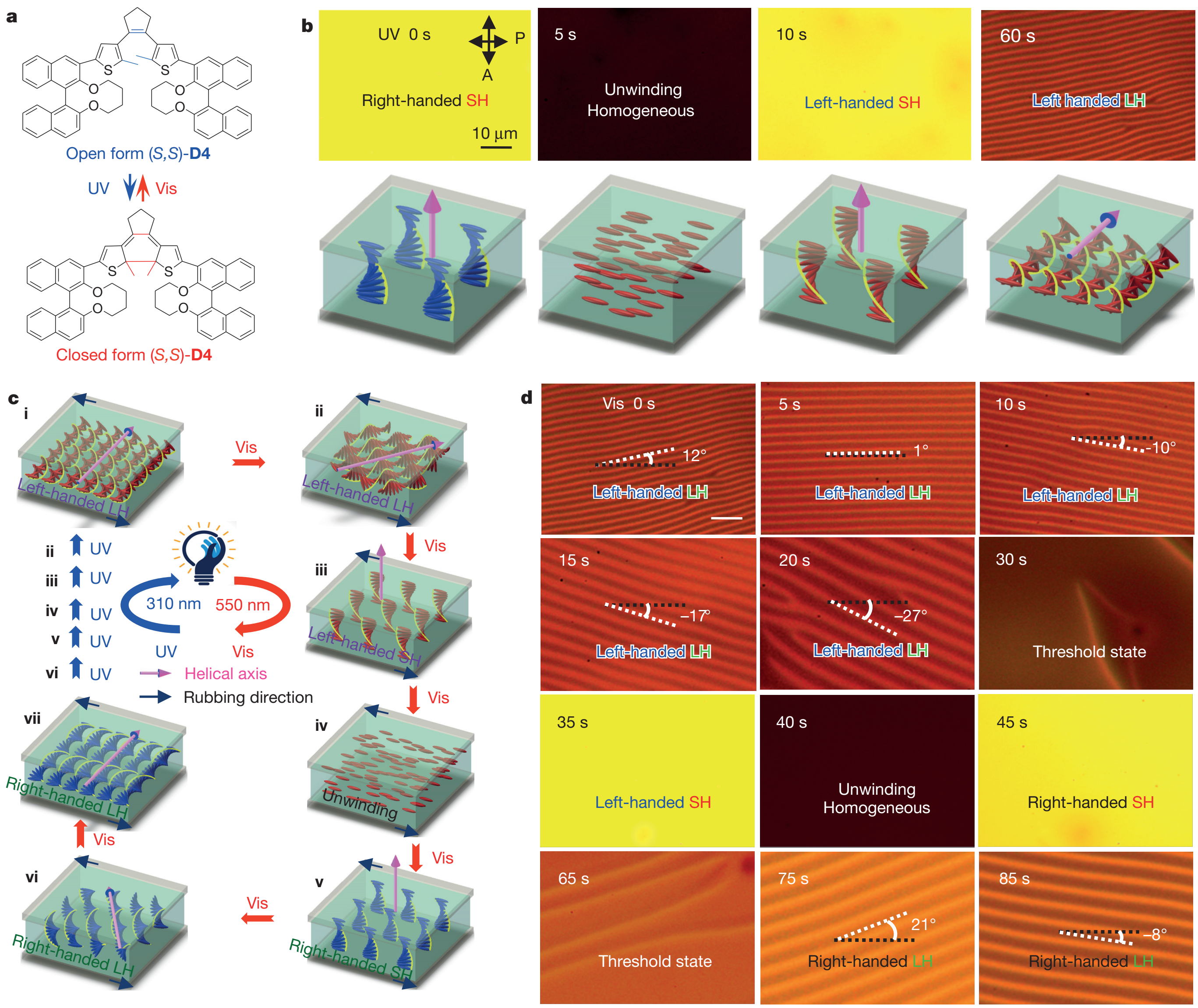


Figure 1 | Light-induced three-dimensional control over the helical axis of a CLC. a, Molecular structure of the photodynamic, switchable, chiral molecular switch with thermal stability, (*S*,*S*)-**D4**. **b**, Ultraviolet (UV) irradiation (wavelength 310 nm, intensity 2.2 mW cm^{-2}) of the chiral switch in a planar cell transforms the helical superstructure from its original standing helices (SH), through the unwound homogeneous state and the SH arrangement with opposite handedness, to the lying helix (LH) arrangement. The LC arrangements are shown schematically below the polarizing optical micrographs. The pink arrows denote the direction of the helical axis in the cell. **c**, Illustration of reversible, light-induced, three-dimensional control over the direction of the helical axis. After the left-handed LH is obtained by UV exposure (**i**), the sample is triggered by visible light (vis; 550 nm)—producing, in sequence, a clockwise in-plane rotation (**ii**); transformation from the left-handed LH to left-handed SH organization (**iii**); unwinding of the left-handed SH to generate a homogeneous alignment (**iv**); and reappearance of the right-handed SH arrangement (**v**). Further stimulation with visible light causes the right-handed standing helices to lie down again (**vi**), and then to form the right-handed lying helices and to rotate clockwise in-plane (**vii**) when the system reaches the visible photostationary state. This whole process can then be driven backwards to the original state by UV light irradiation (blue arrows). This reversible sequence of events in a continuous process establishes the three-dimensional manipulation of the helical axis. **d**, Evolution of the optical texture of the CLC during visible light irradiation from the UV photostationary state. The angles between the direction of the stripes (white dashed lines) and the horizontal line (rubbing direction, black dashed lines) are labelled. The threshold state is defined as the state in which stripes begin to appear from the SH arrangement. All micrographs were taken under polarizing optical microscope with crossed polarizers (P, polarizer; A, analyser).

(achieved by planar surface anchoring of the cell). To investigate these possibilities, we carried out molecular-dynamics simulations of the photoresponsive CLC (Supplementary Fig. 4); the results are consistent with Landau-de Gennes' elastic theory[15]. Specifically, the results indicate an oblique alignment of LC molecules, resulting from the coupling of the elastic energy with the molecular interactions between the chiral switch and LC molecules during photoisomerization. The cell gap-to-pitch ratio (d/P) is another critical factor in LH formation, and represents the coupling effects from the surface anchoring and the twist elastic energy. We found that the measured value of d/P was very close to integer multiples of 0.5, implying that the LH is obtained only when an appropriate trade-off is reached between the surface anchoring and twist elastic energy. The propensity to form a uniform LH arrangement decreased as the d/P value increased.

The direction of the helical axis in the LH arrangement is determined by the azimuthal angle of the director of LC molecules in the middle layer of the cell[26]; changes in this angle lead to the light-induced in-plane rotation of the helical axis. After photoisomerization, the chiral switch undergoes a dramatic change in its molecular structure (Supplementary Fig. 7), which would cause a large change in the LC

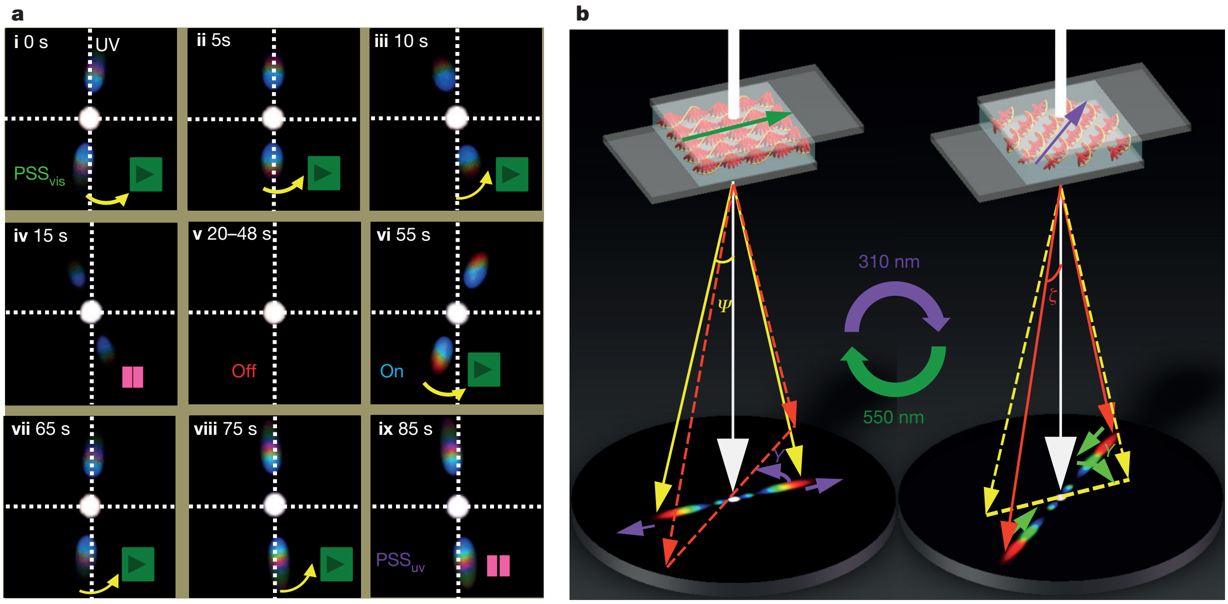


Figure 2 | Light-controllable two-dimensional beam steering for spectrum scanning. a, **i–ix**, Two-dimensional beam steering, resulting from the simultaneous in-plane rotation and pitch modulation of a CLC helix driven by UV light (310 nm). Such beam steering can be turned off (**v**) and on (**vi**) solely by light irradiation. **b**, Diagram showing light-driven, two-dimensional beam steering for spectrum scanning. From left to right, the helical pitch is compressed and the helical axis is rotated during UV irradiation. The pitch compression leads to an increase in the diffraction angle for every wavelength; thus the diffraction pattern shifts away from the centre (purple straight arrow), leading to a change in the diffraction angle at the edge of the pattern (from ψ to ζ). Because of the rotation of the helical axis, the diffraction pattern also rotates anticlockwise (with an angle of γ). The arrows on the LC cells at the top show the direction of the helical axis. The scanning range is defined as $\gamma \times (\zeta - \psi)$. The diffraction pattern can be shortened and rotates clockwise by changing the stimulus to visible light (550 nm). The scanning rate can be further accelerated by increasing the light intensity, and the beam-steering process can be reversibly driven back-and-forth from any intermediate state by alternating UV and visible light irradiation (Supplementary Fig. 14).

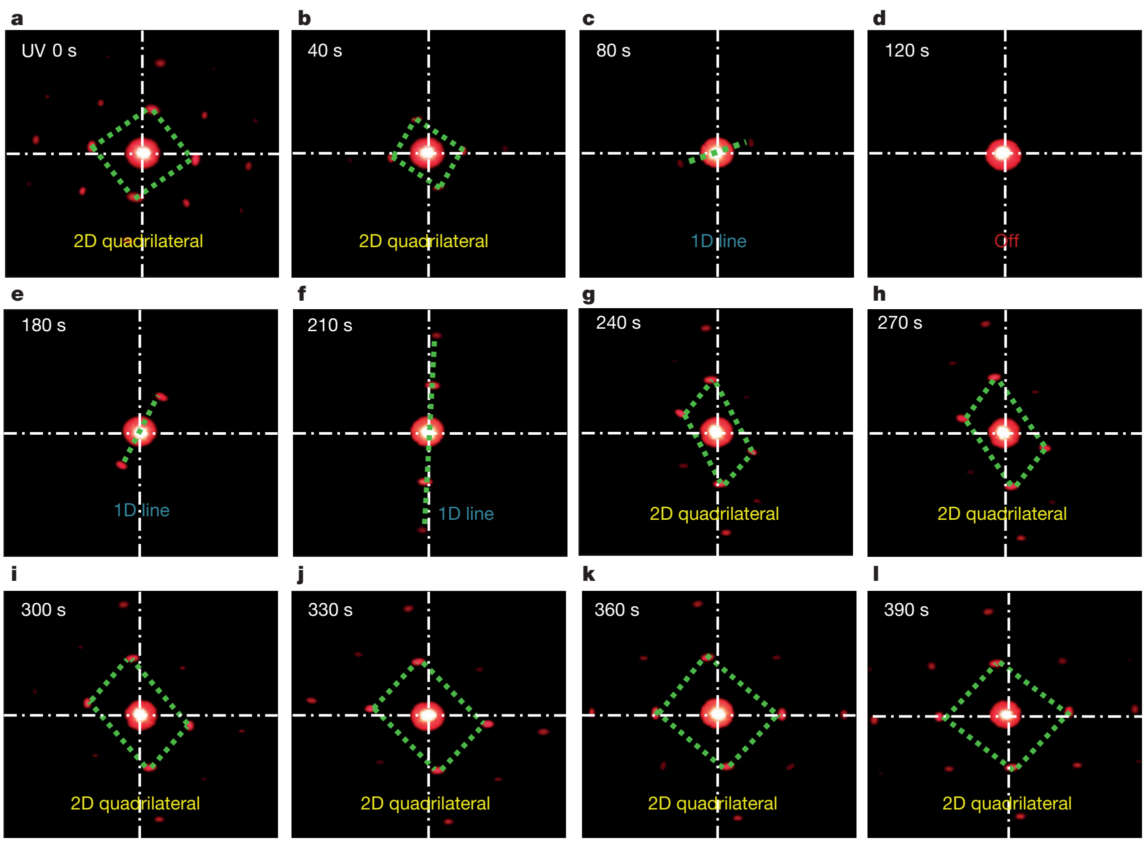


Figure 3 | Light-induced diffraction dimensionality transformation of a bilayer CLC sample. The UV intensity was 0.4 mW cm^{-2}; the irradiation time is shown in the top left of each panel. **a–l**, Diffraction patterns were transformed from two-dimensional (2D; **a**, **b**), through one-dimensional (1D; **c**) and the diffraction off state (**d**), to 1D again (**e**, **f**), and finally to 2D (**g–l**). Green dashed lines, connecting the diffraction spots, reflect the deformation of diffraction properties. The interior angle of the rhomboidal quadrilateral is modulated by rotation of the lying helices within each layer, and the transmitted diffraction angle is modulated by changes in pitch length.

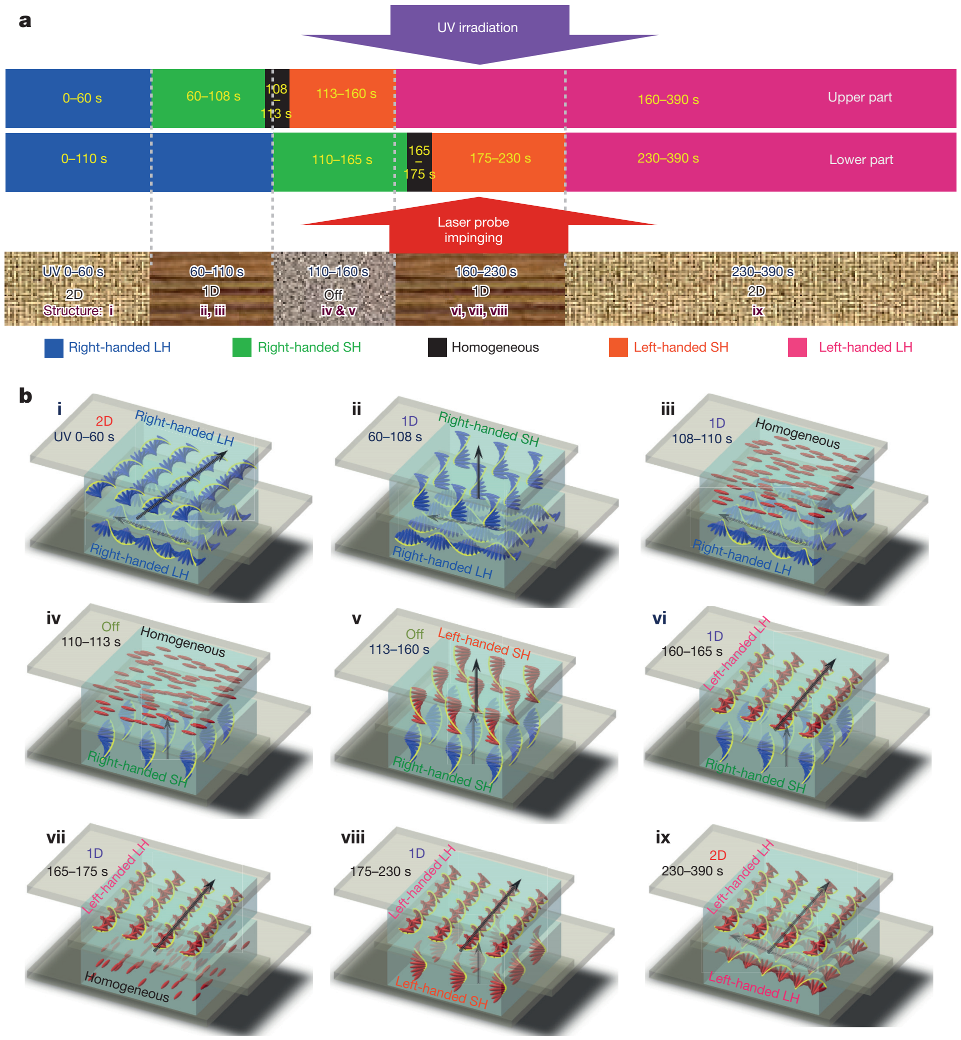


Figure 4 | Helical arrangements of CLCs in a bilayer cell upon UV irradiation. a, Irradiation-time-dependent CLC arrangements in the upper and lower layers of the bilayer (UV intensity, 0.4 mW cm^{-2}). As the UV light propagates from top to bottom, helical transformations occur as shown by the colour key. Combining the arrangements in the upper and lower parts leads to different diffraction effects as a laser probe impinges from bottom to top. Accordingly, there are nine different possible LC arrangements in a bilayer cell, which form different diffraction patterns. **b**, **i–ix**, The nine different possible arrangements, with corresponding time intervals. A one-to-one correspondence between the diffraction pattern and the LC arrangement can be established; for example, the diffraction patterns shown in Fig. 3a (UV 0 s) and Fig. 3b (40 s) correspond to the arrangement shown in panel **b, i** of this figure (UV 0–60 s). The whole process can be briefly described as follows: the 2D diffraction pattern is caused by double effects from two adjacent LH layers (**i**, **ix**); the 1D diffraction pattern results from a bilayer containing an LH layer and either an SH or a homogeneous layer (**ii**, **iii**, **vi**, **vii**, **viii**); the diffraction pattern is switched off if there is no LH arrangement in the bilayer cell (**iv**, **v**).

direction in the middle layer, leading to substantial in-plane rotation of the helical axis (Supplementary Fig. 10). However, the rotation of the helical axis can be suppressed when the LC direction is strongly pinned using an applied electric field (Supplementary Fig. 11). If the surface anchoring is too weak to resist the external disturbance, the formation of the uniform LH arrangement appears not to be favourable (for example, in the case of the larger cell gap shown in Supplementary Figs 8 and 9), or the conventional polydomain fingerprint texture of the CLC is generated (Supplementary Fig. 5k, l). Planar anchoring with a smaller cell gap seems to be favourable for realizing three-dimensional dynamic photocontrol of the CLC helix. Thus, the three-dimensional manipulation of the helical axis depends on a delicate interplay among cell thickness, surface anchoring, pitch length and external stimuli.

To investigate potential applications of this light-induced, three-dimensional manipulation of the CLC helical axis, we explored non-mechanical two-dimensional (in-plane) beam steering (Fig. 2). We observed a chromatic dispersion as a collimated white probe light impinged on the uniform LH arrangement along the cell normal (Fig. 2a(i)). Stimulation with ultraviolet light led to a simultaneous change in helical pitch and in-plane rotation of the stripes of the LH (rotation of the grating vector)—causing the diffraction angle of every wavelength to vary, and enabling two-dimensional in-plane beam steering, which can potentially be applied in spectrum scanning (Fig. 2b). The chromatic dispersion is gradually eliminated by continuous irradiation, because of the decreasing diffraction angle of every wavelength resulting from elongation of the helical pitch. At a time stamp of 20–48 seconds (Fig. 2a(v)) the diffraction has disappeared, because the uniform LH arrangement has transformed into either the SH structure or the unwound homogeneous alignment. Upon further irradiation (to 55 seconds), the LH arrangement with the opposite handedness re-forms and rotates, diffraction reappears, and the diffraction angle increases gradually owing to compression of the CLC pitch, until the sample reaches the PSS (Fig. 2a(ix)). Overall, a wide two-dimensional scanning range of about 52° × 8° (defined in Fig. 2b) is enabled, which is substantially larger than that of 23° × 3.6° reported recently[20]. Such wide, non-mechanical beam steering is desirable

for free-space optical communication, adaptive-optics systems and phased-array radar.

The manipulation and deformation of a two-dimensional beam-spot array is an interesting and challenging task, although metastable two-dimensional gratings have been encountered by chance[27,28], and an electric-field-induced two-dimensional grating in a cholesteric polymer system has been reported[29]. We achieved a reversible dimensionality transformation—from a stable two-dimensional diffraction pattern, via a one-dimensional pattern, to a diffraction off-state—by irradiating a specially designed, LC bilayer cell containing two thin, stacked LH layers (in which the surface directions of the adjacent layers were perpendicular). Figure 3a shows the initial, two-dimensional diffraction array, caused by successive diffraction from two LH layers. When one LH arrangement is converted to either an SH or a homogeneous alignment through irradiation, the diffraction pattern converts from a two-dimensional grid (the rhomboidal quadrilateral encircled by green dashed lines in Fig. 3), to a one-dimensional line, to a direct transmission pattern (indicating the diffraction off-state; Fig. 3d). Further exposure leads to handedness inversion and thus to a reappearance of the LH structures, yielding first a one-dimensional diffraction pattern, and finally the two-dimensional pattern. The deformation of the envelope area and of the interior angles of the rhomboidal quadrilateral arises from the photomodulation of the CLC pitch and in-plane rotation of the helical axis. The time sequence of these changes is due to a progressive fall in the intensity of light, occurring because of photo-absorption by the chiral switch when light passes through the bilayer cell.

The conventional one-dimensional diffraction pattern emanates from one uniform LH layer of the bilayer cell, whereas the two-dimensional diffraction pattern develops as a result of combined diffraction effects from two adjacent LH layers (Fig. 4b(i)). Note that the initial two-dimensional grating (Fig. 3a) arises from two right-handed LH layers, whereas the reappeared two-dimensional diffraction pattern (Fig. 3g–l) results from two left-handed uniform LH layers. The irradiation-time-dependent CLC arrangements in the upper and lower layers are illustrated in Fig. 4a. Figure 4b depicts LC arrangements that might yield the diffraction patterns shown in Fig. 3. It is also conceivable that the one-dimensional diffraction pattern might be switched on or off by changing the incident direction of the probe laser—analogous to the effect of a diode on current—which might enable new optical devices.

In conclusion, we have achieved light-induced, three-dimensional control of the helical axis of self-organized CLCs, resulting in a reversible transformation between an SH and a uniform LH arrangement, with control of both the in-plane rotation angle of the helical axis and the pitch length. This enables reversible, light-driven, wide-area, two-dimensional in-plane beam steering. Moreover, we have accomplished a light-induced reversible transformation between two-dimensional and one-dimensional diffraction patterns and a diffraction off-state by irradiating a bilayer LC cell. The absence of thermal relaxation for this chiral switch enables on-demand, digital control of both the beam direction and the dimensionality of the diffraction array, starting from any desired state and effected exclusively by light. Our work is a step towards the realization of complex, light-activated smart systems and dynamic, reconfigurable three-dimensional architectures.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 October 2015; accepted 18 January 2016.
Published online 7 March 2016.

1. Crissey, J. T., Fergason, J. L. & Bettenhausen, J. M. Cutaneous thermography with liquid crystals. *J. Invest. Dermatol.* **45,** 329–333 (1965).
2. Wu, S. T. & Yang, D. K. *Reflective Liquid Crystal Displays* (John Wiley, 2001).
3. Xiang, J. *et al.* Electrically tunable selective reflection of light from ultraviolet to visible and infrared by heliconical cholesterics. *Adv. Mater.* **27,** 3013–3017 (2015).
4. Ha, N. Y. *et al.* Fabrication of a simultaneous red-green-blue reflector using single-pitched cholesteric liquid crystals. *Nature Mater.* **7,** 43–47 (2008).
5. Coles, H. J. & Morris, S. M. Liquid-crystal lasers. *Nature Photon.* **4,** 676–685 (2010).
6. Chen, L. *et al.* Photoresponsive monodisperse cholesteric liquid crystalline microshells for tunable omnidirectional lasing enabled by a visible light-driven chiral molecular switch. *Adv. Opt. Mater.* **2,** 845–848 (2014).
7. Sackmann, E. Photochemically induced reversible color changes in cholesteric liquid crystals. *J. Am. Chem. Soc.* **93,** 7088–7090 (1971).
8. Bisoyi, H. K. & Li, Q. Light-directing chiral liquid crystal nanostructures: from 1D to 3D. *Acc. Chem. Res.* **47,** 3184–3195 (2014).
9. Subacius, D., Shiyanovskii, S. V., Bos, P. & Lavrentovich, O. D. Cholesteric gratings with field-controlled period. *Appl. Phys. Lett.* **71,** 3323–3325 (1997).
10. Gvozdovskyy, I., Yaroshchuk, O., Serbina, M. & Yamaguchi, R. Photoinduced helical inversion in cholesteric liquid crystal cells with homeotropic anchoring. *Opt. Express* **20,** 3499–3508 (2012).
11. Ponti, S., Ziherl, P., Ferrero, C. & Zumer, S. Flexoelectro-optic effect in a hybrid nematic liquid crystal cell. *Liq. Cryst.* **26,** 1171–1177 (1999).
12. Carbone, G. *et al.* Short pitch cholesteric electro-optical device based on periodic polymer structures. *Appl. Phys. Lett.* **95,** 011102 (2009).
13. Hegde, G. & Komitov, L. Periodic anchoring condition for alignment of a short pitch cholesteric liquid crystal in uniform lying helix texture. *Appl. Phys. Lett.* **96,** 113503 (2010).
14. Outram, B. I., Elston, S. J., Tuffin, R., Siemianowski, S. & Snow, B. The use of mould-templated surface structures for high-quality uniform-lying-helix liquid-crystal alignment. *J. Appl. Phys.* **113,** 213111 (2013).
15. Zola, R. S., Evangelista, L. R., Yang, Y.-C. & Yang, D.-K. Surface induced phase separation and pattern formation at the isotropic interface in chiral nematic liquid crystals. *Phys. Rev. Lett.* **110,** 057801 (2013).
16. Ryabchun, A., Bobrovsky, A., Stumpe, J. & Shibaev, V. Rotatable diffraction gratings based on cholesteric liquid crystals with phototunable helix pitch. *Adv. Opt. Mater.* **3,** 1273–1279 (2015).
17. Patel, J. S. & Meyer, R. B. Flexoelectric electro-optics of a cholesteric liquid crystal. *Phys. Rev. Lett.* **58,** 1538–1540 (1987).
18. Kang, S. W., Sprunt, S. & Chien, L. C. Structure and morphology of polymer-stabilized cholesteric diffraction gratings. *Appl. Phys. Lett.* **76,** 3516–3518 (2000).
19. Kim, S. H., Chien, L. C. & Komitov, L. Short pitch cholesteric electro-optical device stabilized by nonuniform polymer network. *Appl. Phys. Lett.* **86,** 161118 (2005).
20. Hulme, J. C. *et al.* Fully integrated hybrid silicon two dimensional beam scanner. *Opt. Express* **23,** 5861–5874 (2015).
21. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S. & Watts, M. R. Large-scale nanophotonic phased array. *Nature* **493,** 195–199 (2013).
22. Eelkema, R. *et al.* Molecular machines: nanomotor rotates microscale objects. *Nature* **440,** 163 (2006).
23. Jau, H. C. *et al.* Light-driven wide-range nonmechanical beam steering and spectrum scanning based on a self-organized liquid crystal grating enabled by a chiral molecular switch. *Adv. Opt. Mater.* **3,** 166–170 (2015).
24. Li, Y., Xue, C., Wang, M., Urbas, A. & Li, Q. Photodynamic chiral molecular switches with thermal stability: from reflection wavelength tuning to handedness inversion of self-organized helical superstructures. *Angew. Chem. Int. Edn* **52,** 13703–13707 (2013).
25. Wang, L. *et al.* Luminescence-driven reversible handedness inversion of self-organized helical superstructures enabled by a novel near-infrared light nanotransducer. *Adv. Mater.* **27,** 2065–2069 (2015).
26. Lin, C. H., Chiang, R. H., Liu, S. H., Kuo, C. T. & Huang, C. Y. Rotatable diffractive gratings based on hybrid-aligned cholesteric liquid crystals. *Opt. Express* **20,** 26837–26844 (2012).
27. Yeh, H. C., Chen, G. H., Lee, C. R. & Mo, T. S. Photoinduced two-dimensional gratings based on dye-doped cholesteric liquid crystal films. *J. Chem. Phys.* **127,** 141105 (2007).
28. Hrozhyk, U. A., Serak, S. V., Tabiryan, N. V. & Bunning, T. J. Periodic structures generated by light in chiral liquid crystals. *Opt. Express* **15,** 9273–9280 (2007).
29. Ryabchun, A., Bobrovsky, A., Stumpe, J. & Shibaev, V. Electroinduced diffraction gratings in cholesteric polymer with phototunable helix pitch. *Adv. Opt. Mater.* **3,** 1462–1469 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements Q.L. acknowledges support from the Air Force Office of Scientific Research (AFOSR; grant no. FA9950-09-1-0193) and the Air Force Research Laboratory. Z.Z. acknowledges receipt of a Scholarship supported by the China Scholarship Council. T.J.B. acknowledges support from the Materials and Manufacturing Directorate and the AFOSR.

Author Contributions Q.L. and T.J.B. designed the research; Z.Z. carried out the experiments; Y.L. synthesized the chiral dopant; Q.L., Z.Z. and H.K.B. prepared the manuscript; Z.Z, Y.L., H.K.B., L.W., T.J.B. and Q.L. interpreted the results and contributed to manuscript editing.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Q.L. (qli1@kent.edu).

METHODS

The dithienylcyclopentene-based axially chiral molecular switch was synthesized and identified according to the route given in the Supplementary Information, Section 1. The chiral switch was mixed with the commercially available nematic LC E7 at a concentration of 1.2 mol%. The mixture was injected into a planar cell with antiparallel alignment of the substrate surfaces at the clearing point by capillary action, slowly cooled to 52.5 °C, and stored in the dark for about 3 hours to eliminate convection of liquid crystals owing to temperature gradients.

The sample was irradiated using a wide-spectral-band white source (Asahi Spectra LAX-Cute, Japan), fixed with an appropriate narrow-band filter (full width at half-maximum = 10 nm). The intensity of the incoming light was measured using an optical power meter with a highly sensitive silicon-photodiode detector (Thorlabs PM100USB). A fibre-connected white-light source (Ocean Optics LS-1) with a spectral range from 330 nm to 830 nm was used to characterize the performance of the two-dimensional beam steering for spectrum scanning, while a monochromatic laser with a wavelength of 650 nm was collimated and adopted as the probe beam to study the dimensionality transformations of the bilayer sample. A spatial filter (Newport 900PH-25) was used to eliminate the stray light. To achieve two-dimensional beam steering and the bilayer diffraction device, the uniform LH arrangement was first induced by ultraviolet irradiation, then followed by visible light irradiation to drive the sample to PSS_{vis}. The texture transformation during the light irradiation was confirmed by polarizing optical microscopy (Leitz, Laborlux 12 POL), and simultaneously the temperature of the sample was maintained by a hot stage (Mettler Toledo FP82HT). The pitch of the sample was determined through the conventional Cano wedge method.

The bilayer cell was fabricated using three optically transparent flexible polymethyl methacrylate (PMMA) substrates separated by mylar film spacers of identical thickness (4 μm). The substrates were first coated with an antireflection film to eliminate Fresnel reflection and then a polyvinyl alcohol (PVA) film to induce planar homogeneous alignment of the liquid-crystal molecules. The rubbing directions of the PVA films in the upper and lower layers of the bilayer cell were crossed with each other. The structure of the bilayer cell is described in Supplementary Information, Section 12.

LETTER

doi:10.1038/nature17165

The contribution of China's emissions to global climate forcing

Bengang Li[1,2], Thomas Gasser[3,4], Philippe Ciais[3], Shilong Piao[1,5], Shu Tao[1], Yves Balkanski[3], Didier Hauglustaine[3], Juan-Pablo Boisier[3], Zhuo Chen[1], Mengtian Huang[1], Laurent Zhaoxin Li[6], Yue Li[1], Hongyan Liu[1], Junfeng Liu[1], Shushi Peng[1], Zehao Shen[1], Zhenzhong Sun[1], Rong Wang[3], Tao Wang[3], Guodong Yin[1], Yi Yin[3], Hui Zeng[1], Zhenzhong Zeng[1] & Feng Zhou[1]

Knowledge of the contribution that individual countries have made to global radiative forcing is important to the implementation of the agreement on "common but differentiated responsibilities" reached by the United Nations Framework Convention on Climate Change. Over the past three decades, China has experienced rapid economic development[1], accompanied by increased emission of greenhouse gases, ozone precursors and aerosols[2,3], but the magnitude of the associated radiative forcing has remained unclear. Here we use a global coupled biogeochemistry–climate model[4,5] and a chemistry and transport model[6] to quantify China's present-day contribution to global radiative forcing due to well-mixed greenhouse gases, short-lived atmospheric climate forcers and land-use-induced regional surface albedo changes. We find that China contributes 10% $\pm$ 4% of the current global radiative forcing. China's relative contribution to the positive (warming) component of global radiative forcing, mainly induced by well-mixed greenhouse gases and black carbon aerosols, is 12% $\pm$ 2%. Its relative contribution to the negative (cooling) component is 15% $\pm$ 6%, dominated by the effect of sulfate and nitrate aerosols. China's strongest contributions are 0.16 $\pm$ 0.02 watts per square metre for CO_2 from fossil fuel burning, 0.13 $\pm$ 0.05 watts per square metre for CH_4, -0.11 ± 0.05 watts per square metre for sulfate aerosols, and 0.09 $\pm$ 0.06 watts per square metre for black carbon aerosols. China's eventual goal of improving air quality will result in changes in radiative forcing in the coming years: a reduction of sulfur dioxide emissions would drive a faster future warming, unless offset by larger reductions of radiative forcing from well-mixed greenhouse gases and black carbon.

We first estimate the relative contributions of China to the various present-day radiative forcing (RF) components. These relative contributions are then combined with the Intergovernmental Panel on Climate Change (IPCC)'s best guesses[7] of global RFs to deduce China's absolute contributions. Only three known anthropogenic perturbations of RF are missing from our quantitative assessment: secondary organic aerosols, black carbon deposition on snow, and the indirect effects of aerosols.

We apply a new global coupled biogeochemistry–climate model to calculate net changes in atmospheric concentration of well-mixed greenhouse gases (WMGHGs) by balancing prescribed historical anthropogenic emissions against natural removal processes. This model of reduced complexity, called OSCAR v2.1, is enabled by the more complex Earth system models[4,5] upon which its parameters were calibrated. Additionally, the fully fledged LMDz4-INCA3 chemistry and transport model[6] is used to calculate the evolution of the direct RF of short-lived atmospheric climate forcers (SLCFs). Finally, the RF resulting from altered surface albedo following historical land-use change (LUC) in China is estimated through an observation-based surface albedo reconstruction[8].

In all these model integrations, we isolate the marginal contribution of Chinese emissions to the simulated RFs using factorial simulations. This attribution method has been used in previous studies[4,9,10]; and it is further described in Methods. This approach 'tags' each radiatively active species in the models, following the cause-to-effect chain from the emission to the induced radiative forcing. The principle of this causal attribution is in line with the "common but differentiated responsibilities" acknowledged by the parties to the United Nations Framework Convention on Climate Change (UNFCCC)[11,12]. The year 2010 is taken to be the present day, and up until the year 1750 is considered to be pre-industrial, according to the IPCC.

Figure 1 shows the relative and absolute contributions of historical Chinese emissions since 1750 to each component of global RF in 2010. Overall, China contributes 10% $\pm$ 4% (0.30 ± 0.11 W m^{-2} out of 2.88 ± 0.46 W m^{-2}) of the current net global RF from anthropogenic emissions since 1750. This contribution is the sum of two terms with opposite signs. China contributes 12% $\pm$ 2% (0.48 ± 0.09 W m^{-2} out of 4.13 ± 0.40 W m^{-2}) of the global positive RF from WMGHGs, tropospheric ozone and black carbon aerosols, and 15% $\pm$ 6% (-0.18 ± 0.06 out of -1.26 ± 0.24 W m^{-2}) of the global negative RF from LUC-induced surface albedo changes, stratospheric ozone, the effect of ozone precursors on CH_4 lifetime, and sulfate, nitrate and particulate organic matter aerosols.

In absolute terms, the greatest contribution Chinese emissions make to global RF is from CO_2 produced by fossil fuel burning and cement production (0.16 ± 0.02 W m^{-2}). This contribution is followed by that of CH_4 from anthropogenic sources, including its effects on ozone and water vapour (0.13 ± 0.05 W m^{-2}), sulfate aerosols produced by SO_2 emission (-0.11 ± 0.05 W m^{-2}) and black carbon aerosol emission (0.09 ± 0.06 W m^{-2}). In relative terms, China contributes substantially more to the global RF induced by aerosols (for example, 28% $\pm$ 10% of sulfate aerosols, 24% $\pm$ 8% of nitrate aerosols, and 14% $\pm$ 6% of black carbon aerosols) than to the RF induced by WMGHGs.

It is important to note that the contribution of China to current global annual anthropogenic emissions is larger than its contribution to radiative forcings (Fig. 1). For WMGHGs that have long atmospheric lifetimes[7] (from a few decades to several centuries), the legacy of past emissions from countries that began to emit early (such as Europe and the USA) still have a contribution to present-day RF larger than that of China, despite China's much higher emissions nowadays. In contrast, because SLCFs have short atmospheric lifetimes[7] (from days to months), it is the spatial distribution of current emissions, and the local processes controlling their atmospheric transport and

[1]Sino-French Institute for Earth System Science, Laboratory for Earth Surface Processes, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China. [2]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, 210023, China. [3]Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, 91191 Gif-sur-Yvette, France. [4]Centre International de Recherche en Environnement et Développement, CNRS-PontsParisTech-EHESS-AgroParisTech-CIRAD, 94736 Nogent-sur-Marne, France. [5]Key Laboratory of Alpine Ecology and Biodiversity, Institute of Tibetan Plateau Research, Center for Excellence in Tibetan Earth Science, Chinese Academy of Sciences, Beijing 100085, China. [6]Laboratoire de Météorologie Dynamique, CNRS, Université Pierre et Marie Curie—Paris 6, 75252 Paris, France.

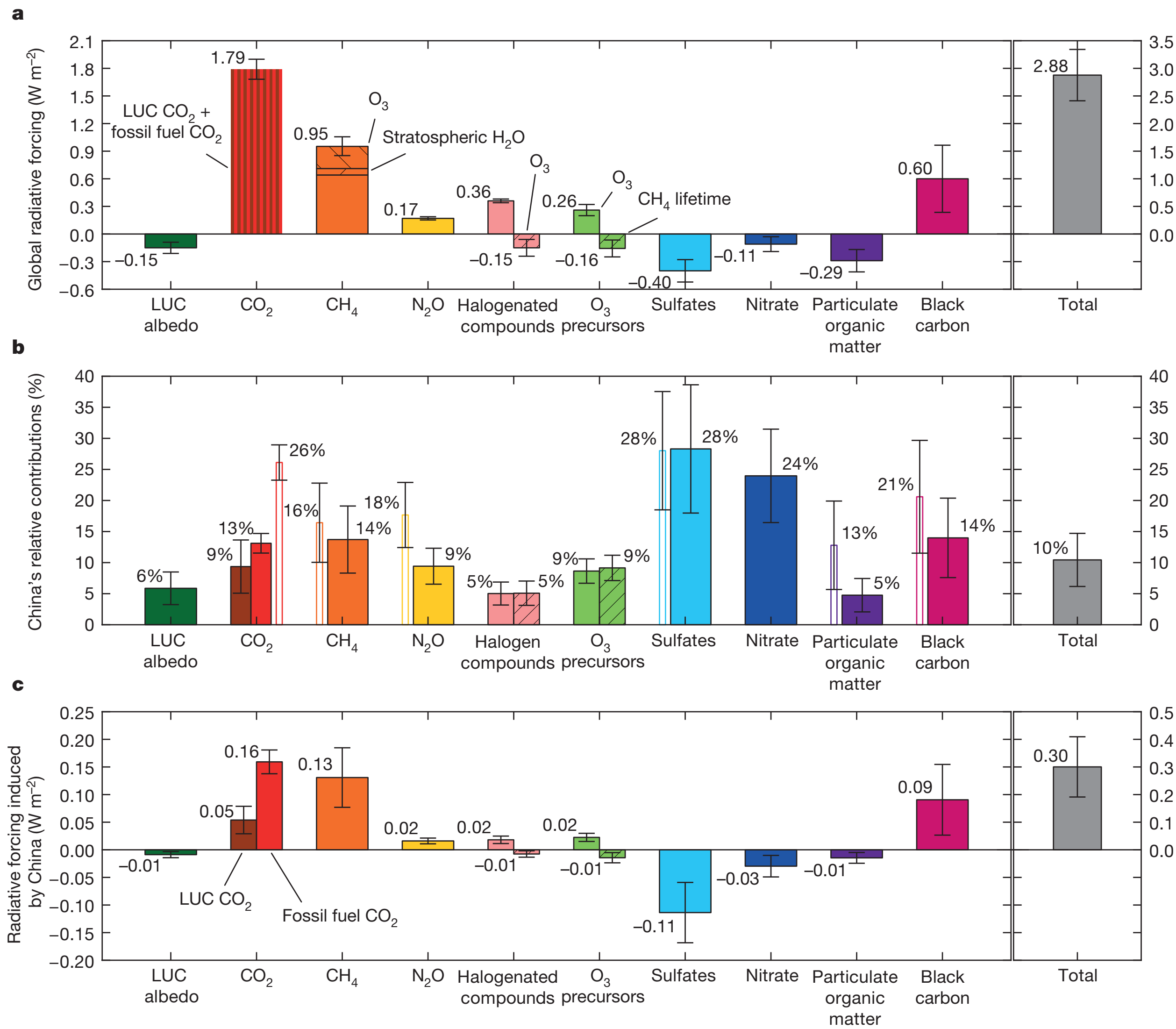


Figure 1 | Attribution of present-day global RF and its components to China. a, The global RF components and their uncertainty, as estimated by the IPCC[7]. **b**, The relative contributions of China to the various components of global RF in 2010, with our assessment of uncertainties (see Methods). When one component of the RF is driven by only one species, China's relative contribution to present-day emissions of that species is also shown as an empty bar oulined in the same colour. **c**, The absolute contributions of China to the RF components and uncertainties, as obtained by combining the values of the two previous panels using a Monte Carlo approach ($n = 50,000$). The 'Total' columns of **a** and **c** are obtained through Monte Carlo summation ($n = 50,000$) of the corresponding RF components; the 'Total' column of **b** is then deduced through the element-wise ratio of these two Monte Carlo ensembles. All uncertainties are one standard deviation.

removal, that determine the contribution of Chinese emissions of ozone precursors and aerosols to the global RF.

Figure 2 shows the absolute and relative historical contributions of Chinese emissions of WMGHGs. During the last 150 years, these contributions were dominated by LUC CO_2 emission. The absolute contribution of China to global RF through its LUC CO_2 emission progressively increased until 1980. After 1980, the LUC CO_2 contribution started decreasing, as the effect of large-scale reforestation and afforestation programmes in China changed the land-use sector into a net sink of CO_2[13–15]. Despite the sharp rise of fossil fuel emissions since the 1960s, it was only around 1992 that the absolute contribution of fossil fuel CO_2 emissions to the RF from China exceeded that of LUC CO_2 emissions. Since that time, fossil fuel CO_2 emissions have been the predominant warming factor. At present, LUC CO_2 remaining in the atmosphere from past emission still accounts for one-third of China's contribution to global RF induced by CO_2, against two-thirds for fossil fuel CO_2. This partitioning occurs despite the fact that the 'instantaneous' fossil fuel emissions of China are today (in 2010) a source of 2.32 ± 0.21 petagrams of carbon per year (Pg C yr^{-1}) ($26\% \pm 3\%$ of the global total), and LUC is a net sink of -0.11 ± 0.04 Pg C yr^{-1} ($-16\% \pm 7\%$ of the global total). The RFs induced by the other WMGHGs emitted by China also show an increasing trend since 1750, but with an acceleration after the 1950s for CH_4 (the absolute contribution of this greenhouse gas to China's RF increased by a factor of about three between 1950 and 2010) and another at the beginning of the 1990s for halogenated compounds.

Despite the increasing absolute contribution of China to the global RF of all WMGHGs, the relative share of China to this RF has remained surprisingly stable (between 8% and 12%) over the last 150 years (Fig. 2, inset). It decreased slightly until 1900, when industry was first introduced to China. It then progressively increased, reaching a maximum around 1960, mainly owing to high LUC CO_2 emissions. From 1960, the relative contribution of China to the global RF of WMGHGs decreased slightly for more than 30 years until 2000, when it again started increasing. The decreasing relative contribution during 1960–1990 is the result of several factors. This period started with the 'Three-year Natural Disaster' and the 'Great Chinese Famine' (1959–1961); this was followed by China's Cultural Revolution

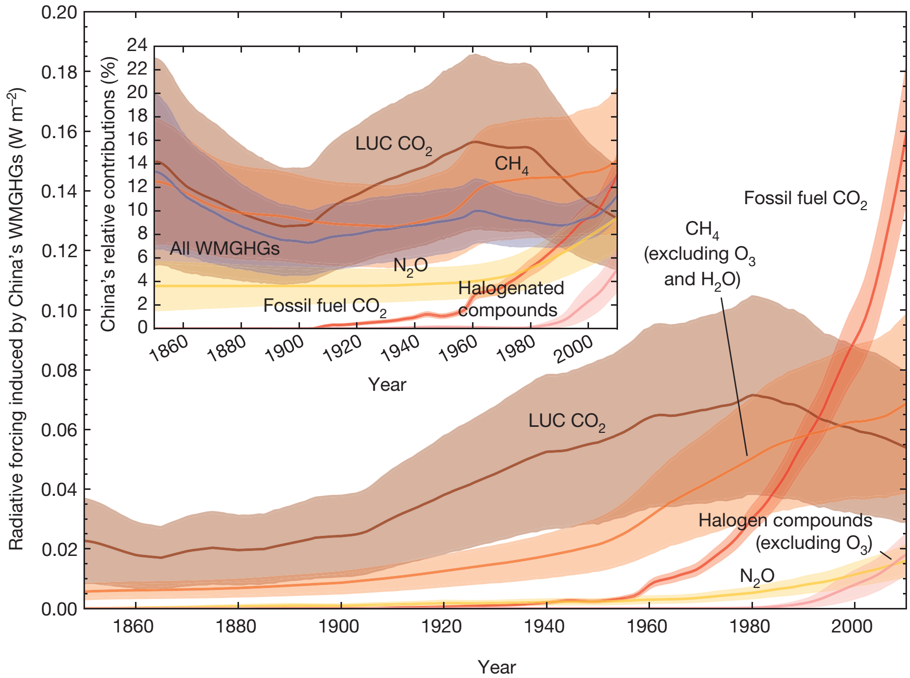


Figure 2 | Historical time series of China's absolute and relative contributions to the RF of WMGHGs. The coloured plain lines are the best-guess absolute or relative contributions of China and the coloured shaded areas are the standard deviations coming from our uncertainty assessment (see Methods). China's relative contributions (inset) are given as the percentage of the global RF of the single compound, except for the 'All WMGHGs' line. See Supplementary Information section A.1 for further discussion on the role of CO_2 from LUC. Contributions are shown only from 1850 for clarity, but pre-industrial is taken as being pre-1750.

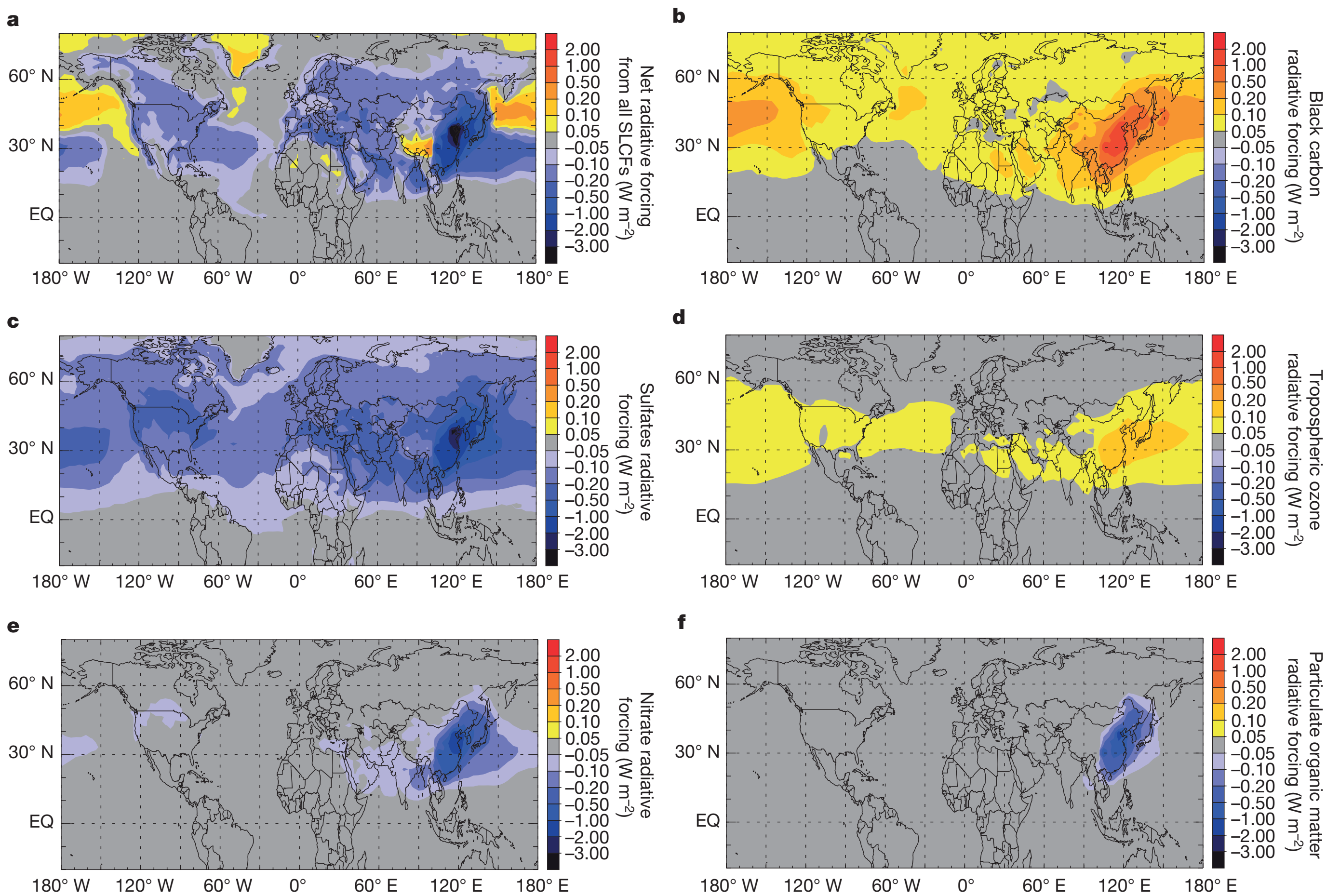


Figure 3 | Spatial distribution of RFs from China-induced SLCFs. All-sky RF of the SLCFs induced by China through emission of short-lived pollutants and precursors in 2010. These RFs are direct outputs from the LMDz-INCA model. **a**, The net RF of all the SLCFs combined. **b**–**f**, The RFs from black carbon, sulfates, tropospheric ozone, nitrate and particulate organic matter, respectively.

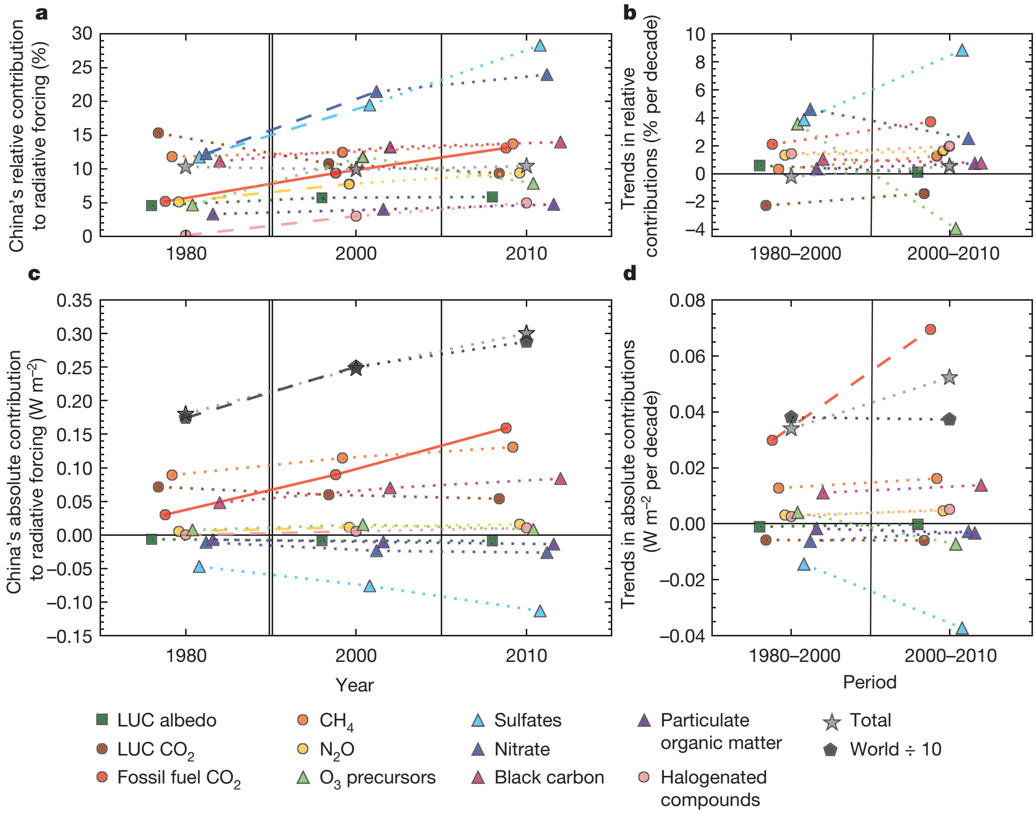


Figure 4 | Recent decadal trends of China's contributions to the global RF and its components. a, The evolution of China's relative contributions over the last thirty years with time slices in 1980, 2000 and 2010. **b,** The decadal trends of these relative contributions. **c, d,** As for **a** and **b**, but for China's absolute contributions. The absolute value of global RF and its trend, divided by a factor of ten, is also shown, in black. Uncertainties are not shown for the sake of readability. However, solid lines mark statistically significant trends ($P < 0.05$) between two periods, and dashed lines mark moderately significant trends ($P < 0.20$). (Dotted lines mark trends that are not significant.) Thus, a solid line in **a** or **c** shows a significant trend in speed of change (first derivative) in China-induced RF, and one in **b** or **d** shows a significant trend in its acceleration (second derivative). Symbols are introduced for clarity only.

(1966–1976), during which industrial activities and emissions slowed down. From 1982 onward, China's reform and open policies led to the current period of rapid economic growth[1], resulting in increased RF of WMGHGs. However, at the same time as industrial activities were developing in the 1980s, changed land-use policy dominated by forest planting resulted in LUC emission declining and eventually creating a carbon sink[13–15] (see also Supplementary Information section A.1). The albedo effect of LUC is discussed in Supplementary Information section A.2.

The global direct RF from China resulting from all SLCFs together is estimated to be $-0.05 \pm 0.09\ \mathrm{W\ m^{-2}}$ in 2010, which corresponds to a slight yet not statistically significant cooling effect. However, this value is the difference between two large and opposite terms: a large negative RF caused by sulfate, nitrate and particulate organic matter aerosols, and a large positive RF caused by ozone precursors and black carbon aerosols (Fig. 1 and Supplementary Fig. 1). It is thus in fact fortuitous that today the cooling and warming effects of the different SLCFs emitted by China nearly compensate for one another.

Note that the direct RF of SLCFs at the regional scale over China is much more pronounced than at the global scale (Fig. 3). The combined RF of SLCFs caused by China is negative over East Asia, and remains slightly negative over most of the globe. For China's SLCF emissions, the RF effect of particulate organic matter is mostly local. The effect of nitrate aerosols influences the whole of East Asia, while the RF of sulfate, black carbon and tropospheric ozone caused by emissions in China affects the entire Northern Hemisphere; this is due to the longer lifetimes of these species in the atmosphere with respect to chemistry and deposition processes, compared to other SLCFs.

Over the last thirty years, China's relative contribution has remained close to 10% of global RF (Fig. 4a). This contribution shows a slight positive trend over the period (less than a percentage point per decade; Fig. 4b), although it is not statistically significant. The absolute contribution of China to global RF has increased over the same period (Fig. 4c) with a noticeable—yet again not significant—acceleration over the last decade happening while the trend of global RF has remained approximately steady (Fig. 4d), thus explaining the slight positive trend in China's relative contribution. As before, these trends result from opposite trends caused by warming and cooling species, as also shown in Fig. 4.

Of all the species considered in this study, only the positive trend in the relative and absolute contributions of China to the RF of fossil fuel CO_2 emissions is statistically significant ($P < 0.05$). This species also exhibits a moderately significant ($P < 0.20$) acceleration of its absolute contribution to global RF, meaning that the strong increase in Chinese emission of fossil fuel CO_2 over the past decades can already be detected, in terms of RF, by our attribution framework. Despite not being statistically significant, the trends we estimate for the RF of other species are worth mentioning because they provide key elements in assessing the near-future contribution of China to global climate forcing. We find a positive trend in Chinese relative and absolute contributions to the RF induced by all WMGHGs except LUC CO_2; and these trends see minor accelerations over the recent past. For SLCFs, China's relative contribution to the global RF of sulfate, nitrate, particulate organic matter and black carbon aerosols has also increased over the last thirty years. Chinese sulfates show an accelerating contribution, whereas nitrate shows a decelerating contribution to the net cooling effect. China's particulate organic matter and black carbon

contributions increased at a relatively steady pace. Finally, the recently decelerating net RF attributed to China's emission of ozone precursors results from two opposite effects: a steadily increasing tropospheric ozone burden (which is a warming effect), and an accelerating reduction of the CH_4 lifetime mainly caused by accelerating NO_x emissions (a cooling effect).

Given the abovementioned estimates of trends, and the delay in reducing emissions caused by the lifetime of existing emitting infrastructure[16] or by social and political systems[17], and given China's pledges in the global climate negotiations (http://unfccc.int/focus/indc_portal/items/8766.php), it appears unlikely that the Chinese contribution to the global RF through greenhouse gas emissions will decrease in the coming years. This alone will increase China's contribution to global RF in the coming years.

However, it is possible that China may choose to improve its air quality by reducing pollutant emissions. This improvement could happen through reduction of sulfur dioxide emissions, which would reduce the current cooling effect of sulfate aerosols (see Supplementary Information section A.3). Whether this would imply an even faster rise of China's contribution to the global RF would ultimately depend on the reduction of other (sometimes co-emitted) pollutants such as black carbon and ozone precursors[18,19].

We found that China's emissions contribute about 10% to the global RF in 2010. Such a percentage is commensurate with the Chinese population, size and economy (19% of the world population, 6% of global emerged land, and 12% of the global GDP in 2012). Our estimates of China's contribution to global RF through its emissions of WMGHGs are consistent with previous results obtained with a simpler model[10]. However, contributions from CH_4 and LUC CO_2 shows high uncertainty, which reveals the need for more accurate emission inventories.

Contributions from all SLCFs also showed high uncertainty, which has two causes. First, knowledge of emissions of those species is limited, especially of sulfur dioxide over China[20,21]. To improve this we need better emission inventories, and improved monitoring networks. Second, limited scientific understanding of aerosol–radiation interactions[22] and aerosol–cloud interactions[22,23] makes quantitative RF estimates of aerosols and SLCFs more uncertain, one reason why the latter interactions are not quantitatively assessed here. The three missing RF components of this study are qualitatively assessed in Supplementary Information section A.4. The inclusion of secondary organic aerosols and black carbon deposition on snow would slightly increase China's relative contribution. Inclusion of the aerosol–cloud interactions, however, would noticeably reduce China's relative contribution to the present-day global RF because of an even stronger masking effect induced by Chinese aerosol emissions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 May 2014; accepted 19 January 2016.

1. National Bureau of Statistics of China. *Gross Domestic Product Index Since 1978* http://data.stats.gov.cn/english/easyquery.htm?cn=C01 (see 'National Accounts' for GDP) (2014).
2. Boden, T. A., Marland, G. & Andres, R. J. *Global, Regional, and National Fossil-Fuel CO_2 Emissions* http://dx.doi.org/10.3334/CDIAC/00001_V2013 (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US DOE, 2013).
3. *Emission Database for Global Atmospheric Research (EDGAR)* Release version 4.2, http://edgar.jrc.ec.europa.eu (European Commission, Joint Research Centre, Netherlands Environmental Assessment Agency, 2011).
4. Gasser, T. *Attribution Régionalisée des Causes Anthropiques du Changement Climatique* https://tel.archives-ouvertes.fr/tel-01135456, PhD thesis, Univ. Pierre et Marie Curie (2014).
5. Cherubini, F., Gasser, T., Bright, R. M., Ciais, P. & Stromman, A. H. Linearity between temperature peak and bioenergy CO_2 emission rates. *Nature Clim. Change* **4,** 983–987 (2014).
6. Hauglustaine, D. A. *et al.* Interactive chemistry in the Laboratoire de Météorologie Dynamique general circulation model: description and background tropospheric chemistry evaluation. *J. Geophys. Res.* **109,** D04314 (2004).
7. Myhre, G. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) 659–740 (Cambridge Univ. Press, 2013).
8. Boisier, J. P., de Noblet-Ducoudré, N. & Ciais, P. Inferring past land use-induced changes in surface albedo from satellite observations: a useful tool to evaluate model simulations. *Biogeosciences* **10,** 1501–1516 (2013).
9. Ciais, P. *et al.* Attributing the increase in atmospheric CO_2 to emitters and absorbers. *Nature Clim. Change* **3,** 926–930 (2013).
10. Höhne, N. *et al.* Contributions of individual countries emissions to climate change and their uncertainty. *Clim. Change* **106,** 359–391 (2011).
11. United Nations Framework Convention on Climate Change (UNFCCC). *Methodological Issues: Scientific and Methodological Assessment of Contributions to Climate Change*, Report of the Expert Meeting, Note by the Secretariat, http://unfccc.int/resource/docs/2002/sbsta/inf14.pdf (UNFCCC, 2002).
12. Trudinger, C. & Enting, I. Comparison of formalisms for attributing responsibility for climate change: non-linearities in the Brazilian proposal approach. *Clim. Change* **68,** 67–99 (2005).
13. Houghton, R. A. & Hackler, J. L. Sources and sinks of carbon from land-use change in China. *Glob. Biogeochem. Cycles* **17,** 1034 (2003).
14. Piao, S. L. *et al.* The carbon balance of terrestrial ecosystems in China. *Nature* **458,** 1009–1013 (2009).
15. Jain, A. K., Meiyappan, P., Song, Y. & House, J. I. CO_2 emissions from land-use change affected more by nitrogen cycle, than by the choice of land-cover data. *Glob. Change Biol.* **19,** 2893–2906 (2013).
16. Davis, S. J., Matthews, D. & Caldeira, K. Future CO_2 emissions and climate change from existing energy infrastructure. *Science* **329,** 1330–1333 (2010).
17. Ha-Duong, M., Grubb, M. J. & Hourcade, J. C. Influence of socioeconomic inertia and uncertainty on optimal CO_2-emission abatement. *Nature* **390,** 270–273 (1997).
18. Boucher, O. & Reddy, M. S. Climate trade-off between black carbon and carbon dioxide emissions. *Energy Policy* **36,** 193–200 (2008).
19. Shindell, D. *et al.* Simultaneously mitigating near-term climate change and improving human health and food security. *Science* **335,** 183–189 (2012).
20. Klimont, Z., Smith, S. J. & Cofala, J. The last decade of global anthropogenic sulphur dioxide: 2000–2011 emissions. *Environ. Res. Lett.* **8,** 014003 (2013).
21. Huneeus, N., Boucher, O. & Chevallier, F. Atmospheric inversion of SO_2 and primary aerosol emissions for the year 2010. *Atmos. Chem. Phys.* **13,** 6555–6573 (2013).
22. Boucher, O. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) 571–657 (Cambridge Univ. Press, 2013).
23. Carslaw, K. S. *et al.* Large contribution of natural aerosols to uncertainty in indirect forcing. *Nature* **503,** 67–71 (2013).
24. Yu, H. *et al.* A multimodel assessment of the influence of regional anthropogenic emission reductions on aerosol direct radiative forcing and the role of intercontinental transport. *J. Geophys. Res.* **118,** 700–720 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Schulz for help with Supplementary Fig. 7, and H. Yu for sharing data from ref. 24. This study is supported by the National Natural Science Foundation of China (grant numbers 41371443, 41390240) and the 111 project (grant number B14001). It is also part of the ACACCYA project funded by the GIS Climat-Environnement-Société. T.G. is supported by the European Research Council Synergy grant ERC-2013-SyG-610028 IMBALANCE-P.

Author Contributions B.L., T.G., P.C., S. Piao and S.T. designed the study. Simulations and output analysis were performed by T.G. for OSCAR and the overall integration; by D.H., R.W. and Y.B. for LMDz-INCA; by J.-P.B. and L.Z.L. for LUC albedo reconstructions; by Y.B. for black carbon albedo; by Z.C. and Y.B. for secondary organic aerosols; and by B.L., T.G., D.H. and R.W. for model evaluation. B.L., S. Peng, Y.Y. and F.Z. provided additional data and analysis. Writing was led by B.L., with substantial inputs from T.G., P.C., S. Piao, S.T., Y.B., D.H. and R.W. All authors participated in the study, the interpretation of the results, and the outline of the paper, through regular meetings and discussion over the past three years.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.L. (lbg@urban.pku.edu.cn).

METHODS

This study's first goal is to assess China's relative contribution to present-day (that is, 2010) RF. Its absolute contribution to present-day RF is then deduced by multiplying our estimates of relative contribution by the IPCC estimate of global RF[7]. Uncertainties are assessed through Monte Carlo ensembles ($n = 50{,}000$).

To calculate China's relative contribution to the RF from WMGHGs, we use a simple model of global biogeochemical cycles named OSCAR v2.1 (refs 4 and 5). It is an update of the version 2.0 (ref. 25) where non-CO_2 gases were added. OSCAR calculates the global concentration of CO_2, CH_4, N_2O and halogenated compounds, such as SF_6, NF_3, many HFCs, PFCs, CFCs, HCFCs (see Supplementary Information) by balancing their historical anthropogenic emissions against the (nonlinear) natural removal processes that define the lifetime of each species. The representation of these natural processes has been calibrated on more complex (usually three-dimensional) models. A complete description of the model is provided in Supplementary Information section C. We used several data sets for the historical emission of WMGHGs, some of which have a specific focus on China[13,26,27], and we combined the corresponding simulations by OSCAR to obtain our best guess of China's relative contribution (see also Supplementary Information section B.1). More details on the performance of OSCAR and the experimental set-up can be found in Supplementary Information sections B.2 and B.3.

To calculate China's relative contribution to the RF from SLCFs, we use a three-dimensional atmospheric-chemistry aerosol model, LMDz4-INCA3 (ref. 6). The RFs of aerosols and ozone are large over the emission region, but decay rapidly away from the region of emission. Hence, it is important that the modelled emissions, transformations and atmospheric removal processes of aerosols and ozone are spatially explicit. The model used in this study is 3.75° in longitude and 1.9° in latitude with 39 levels. A detailed model description can be found in ref. 6. The model is used to estimate China's relative contribution to RF in 1980, 2000 and 2010. Note that for this latter year (2010) we use the most up-to-date emission data set that we know of[28]. Performance of the model in reproducing observed surface concentration of aerosols is shown in Supplementary Fig. 7 and Supplementary Information section B.4.2. No significant bias is found over China or at the global scale. More details regarding the performance of the model and the data sets used are given in Supplementary Information sections B.4 and B.5.

To calculate China's relative contribution to the LUC-induced albedo change and associated RF, we use a diagnostic method that combines present-day satellite data of surface albedo and snow cover with historical land-cover maps[8]. The albedo-induced RF is estimated as a net radiation anomaly at the top of the atmosphere, based on a simple (zero-dimensional) transfer model of shortwave radiation (see Supplementary Information section B.6.1).

To isolate China's relative contribution to the global RF in all those simulations, we use an attribution and linearization method consistent with the Brazilian Proposal (http://www.match-info.net/) and advised by the UNFCCC[11]. The method is called 'normalized marginal' and it attributes to China a relative contribution to RF proportional to the marginal effect of its own emission causing the RF. To calculate China's relative contribution to the global RF, for each WMGHG or SLCF and the LUC albedo, we performed three simulations: one 'normal' with all emissions included in the simulation ('all'); one with Chinese emissions reduced by a fraction ε ('$-\varepsilon_{\mathrm{China}}$'); and one with all except the Chinese emissions reduced by the same fraction ('$-\varepsilon_{\mathrm{RoG}}$', for rest of the globe). We deduce China's relative contribution α following the normalized marginal method: $\alpha = [\mathrm{RF(all)} - \mathrm{RF}(-\varepsilon_{\mathrm{China}})]/[2 \times \mathrm{RF(all)} - \mathrm{RF}(-\varepsilon_{\mathrm{China}}) - \mathrm{RF}(-\varepsilon_{\mathrm{RoG}})]$. Because of model specificities and species properties, different ε values are used: 0.1% with OSCAR, 20% with LMDz-INCA and 100% with the diagnostic method for LUC albedo (corresponding to the 'normalized residual' method). However, several studies have found only small differences in estimated contributions when using different ε values[10,11].

Uncertainties are assessed in terms of the relative contributions of China, which implies that uncertainties in China's absolute contributions are obtained using Monte Carlo ensembles ($n = 50{,}000$) in which we multiply the relative contributions by the global estimates of RF by the IPCC (see Supplementary Information section B.7). We assume that the relative uncertainty in China's relative contribution to one of the components of the global RF is the product of two types of uncertainty. The first is the relative uncertainty in China's share of global emissions (see Supplementary Information section B.3.2 for WMGHGs and Supplementary Information section B.5.2 for SLCFs). It dominates the overall uncertainty. The second is the relative uncertainty in the radiative efficiency of China's emission when compared to globally averaged radiative efficiency (see Supplementary Information section B.3.3 for WMGHGs and Supplementary Information section B.5.3 for SLCFs). For WMGHGs, the difference in radiative efficiency comes from the difference in temporality of emission and the nonlinearity of the system; for SLCFs, it comes from the difference between local and globally averaged processes. For most species, the first type of uncertainty is assessed on the basis of the comparison of multiple data sets[29], and the second type is on the basis of the comparison of multiple models[24,30]. For some species, however, we made assumptions. Details of the uncertainty analysis are given in Supplementary Information section B.3 for WMGHGs, Supplementary Information section B.5 for SLCFs, Supplementary Information section B.6.2 for LUC albedo, and Supplementary Information section B.7 for the overall analysis.

Data availability. Most of the input data used in this study are freely available online: CDIAC (http://cdiac.ornl.gov/trends/emis/meth_reg.html); EDGAR (http://edgar.jrc.ec.europa.eu/overview.php?v=42); EDGAR-FT (http://edgar.jrc.ec.europa.eu/overview.php?v=42FT2010); EDGAR-HTAP (http://edgar.jrc.ec.europa.eu/htap_v2/index.php); ACCMIP (http://tntcat.iiasa.ac.at:8787/RcpDb/dsd?Action=htmlpage&page=download; file 'Historical emissions data (1850–2000)'); EPA (http://www3.epa.gov/climatechange/EPAactivities/economics/nonco2projections.html); EDGAR-HYDE (http://themasites.pbl.nl/tridion/en/themasites/edgar/emission_data/edgar-hyde-100yr/edgar-hyde-1-4.html); PKU-FF (http://inventory.pku.edu.cn/download/download.html); PKU-CH_4 (for China only) (http://dods.extra.cea.fr/work/p24peng/PKU-CH4); RCP ODS emissions (http://www.pik-potsdam.de/~mmalte/rcps/); ECLIPSE (http://eccad.sedoo.fr/eccad_extract_interface/JSF/page_login.jsf); EANET observations (http://www.eanet.asia/index.html); EPA observations (http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html); EMEP observations (http://ebas.nilu.no/); and AeroCom observations (http://aerocom.met.no/cgi-bin/aerocom/surfobs_annualrs.pl).

Code availability. All other input data are provided with the source code of OSCAR. The source code of LMDz-INCA can be downloaded at http://www-lsceinca.cea.fr/. We also provide the source code of OSCAR, used to estimate China's contribution to WMGHGs and to perform the overall integration and uncertainty analysis, along with any input data used in this study as Supplementary Data.

25. Gasser, T. & Ciais, P. A theoretical framework for the net land-to-atmosphere CO_2 flux and its implications in the definition of "emissions from land-use change". *Earth Syst. Dyn.* **4,** 171–186 (2013).
26. Wang, R. *et al.* High-resolution mapping of combustion processes and implications for CO_2 emissions. *Atmos. Chem. Phys.* **13,** 5189–5203 (2013).
27. Zhou, F. *et al.* A new high-resolution N_2O emission inventory for China in 2008. *Environ. Sci. Technol.* **48,** 8538–8547 (2014).
28. Stohl, A. *et al.* Evaluating the climate and air quality impacts of short-lived pollutants. *Atmos. Chem. Phys.* **15,** 10529–10566 (2015).
29. Granier, C. *et al.* Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980–2010 period. *Clim. Change* **109,** 163–190 (2011).
30. Fry, M. M. *et al.* The influence of ozone precursor emissions from four world regions on tropospheric composition and radiative climate forcing. *J. Geophys. Res.* **117,** D07306 (2012).

doi:10.1038/nature17155

Reversal of ocean acidification enhances net coral reef calcification

Rebecca Albright[1], Lilian Caldeira[1], Jessica Hosfelt[2], Lester Kwiatkowski[1], Jana K. Maclaren[1,3], Benjamin M. Mason[4], Yana Nebuchina[1], Aaron Ninokawa[2], Julia Pongratz[1,5], Katharine L. Ricke[1,6], Tanya Rivlin[7,8], Kenneth Schneider[1,9], Marine Sesboüé[1], Kathryn Shamberger[10,11], Jacob Silverman[12], Kennedy Wolfe[13], Kai Zhu[1,14,15] & Ken Caldeira[1]

Approximately one-quarter of the anthropogenic carbon dioxide released into the atmosphere each year is absorbed by the global oceans, causing measurable declines in surface ocean pH, carbonate ion concentration ($[CO_3^{2-}]$), and saturation state of carbonate minerals (Ω)[1]. This process, referred to as ocean acidification, represents a major threat to marine ecosystems, in particular marine calcifiers such as oysters, crabs, and corals. Laboratory and field studies[2,3] have shown that calcification rates of many organisms decrease with declining pH, $[CO_3^{2-}]$, and Ω. Coral reefs are widely regarded as one of the most vulnerable marine ecosystems to ocean acidification, in part because the very architecture of the ecosystem is reliant on carbonate-secreting organisms[4]. Acidification-induced reductions in calcification are projected to shift coral reefs from a state of net accretion to one of net dissolution this century[5]. While retrospective studies show large-scale declines in coral, and community, calcification over recent decades[6–12], determining the contribution of ocean acidification to these changes is difficult, if not impossible, owing to the confounding effects of other environmental factors such as temperature. Here we quantify the net calcification response of a coral reef flat to alkalinity enrichment, and show that, when ocean chemistry is restored closer to pre-industrial conditions, net community calcification increases. In providing results from the first seawater chemistry manipulation experiment of a natural coral reef community, we provide evidence that net community calcification is depressed compared with values expected for pre-industrial conditions, indicating that ocean acidification may already be impairing coral reef growth.

The aragonite saturation state (Ω_{arag}) of tropical surface waters has decreased from about 4.5 in pre-industrial time[13] to approximately 3.8 by 1995 (ref. 14). In this study, sodium hydroxide (NaOH) was used to increase the total alkalinity of seawater flowing over a reef flat, with the aim of increasing $[CO_3^{2-}]$ and Ω_{arag} closer to values that would have been attained under pre-industrial levels of atmospheric CO_2 partial pressure (p_{CO_2}). We used a dual tracer regression method to estimate changes in alkalinity uptake (that is, net community calcification) in response to alkalinity addition. This approach uses the change in ratios between an active tracer (alkalinity) and a passive tracer (a non-reactive dye, Rhodamine WT) to assess the fraction of added alkalinity taken up by the reef. Changes in the active tracer (alkalinity) result from mixing, dilution, and biological activity (that is, calcification), whereas changes in the passive tracer (hereafter referred to as the 'dye') are due solely to mixing and dilution. By comparing the alkalinity to dye ratios before (upstream of the study site) and after (downstream) the water mass interacts with the reef, we were able to isolate the change in alkalinity that is due to an induced increase in net calcification (Extended Data Fig. 1).

Our study was conducted on One Tree Reef (23° 30′ S, 152° 06′ E), a pseudo-atoll in the southern Great Barrier Reef (Fig. 1a). One Tree Reef encloses three lagoons, two of which are hydrologically distinct (that is, separated by reef walls). At low tide, the water level drops below the outer reef crest, and the lagoons are effectively isolated from the ocean (Fig. 1c). Because First Lagoon sits approximately 30 cm higher than Third Lagoon, gravity-driven, unidirectional flow results from First Lagoon, over the reef flat separating the two lagoons, and into Third Lagoon. Our study site was situated along a section of the reef wall separating First and Third Lagoons. Unidirectional flow across this area of the reef flat persists for approximately 60 min following peak low tide, enabling an experimental setup depicted in Fig. 1d. This section of the reef flat is a well-developed, mixed reef community characterized by ~17% live coral (Extended Data Fig. 2).

Our study was conducted once per day, over 22 days between the dates of 16 September 2014 and 10 October 2014. Dates, times, light data, and predicted heights of low tides are provided in Extended Data Table 1. Before low tide each day, a 15 m^3 tank was deployed in First Lagoon, adjacent to the study site. On all 22 days, 4 g Rhodamine WT were mixed with ambient seawater inside the tank. On 15 of those days (hereafter referred to as 'experiment' days), 15 mol (600 g) of NaOH was also introduced into the tank. The resulting solution was pumped onto the reef flat at a constant rate of $\sim 21\,s^{-1}$ for 60 min starting at the predicted time of low tide. The resulting plume flowed over the reef flat as described in the Methods. Following the 60 min pumping period, discrete water samples were taken at pre-defined sampling locations along the length of two parallel transects that defined the borders of the study area (along the upstream and downstream edges of the reef flat; Fig. 1d and Extended Data Fig. 3). Samples were analysed for total alkalinity, rhodamine, pH, dissolved inorganic carbon, and nutrients, as described in the Methods (Supplementary Table 1). On 7 days, observations were made when dye, but no alkalinity, was added (hereafter referred to as 'control days'), to test whether the dye addition had unexpected effects, and to characterize background variability in the study area. Mean chemical conditions for control and experimental days are provided in Fig. 2 and Extended Data Fig. 4. On experiment days, the mean concentration of added alkalinity in the central part of the plume (containing 50% of the dye), was $50.2 \pm 2.7\,\mu mol\,kg^{-1}$, resulting in an average elevation of Ω_{arag} in this part of the plume by 0.6 units. Mean temperatures, salinities, nutrient concentrations,

[1]Department of Global Ecology, Carnegie Institution for Science, Stanford, California 94305, USA. [2]Bodega Marine Laboratory, University of California, Davis, Bodega Bay, California 94923, USA. [3]Stanford Nano Shared Facilities, Stanford University, Stanford, California 94305, USA. [4]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. [5]Max Planck Institute for Meteorology, Bundesstraße 53, 20146 Hamburg, Germany. [6]Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, New York 14853, USA. [7]The Interuniversity Institute for Marine Sciences, The H. Steinitz Marine Biology Laboratory, The Hebrew University of Jerusalem, Eilat, Israel. [8]The Fredy and Nadine Herrman Institute of Earth Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem, Israel. [9]Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel. [10]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. [11]Texas A&M University, College Station, Texas 77843, USA. [12]Institute for Oceanographic and Limnological Research, Haifa, Israel. [13]School of Medical Sciences, The University of Sydney, Sydney, New South Wales 2006, Australia. [14]Department of Biology, Stanford University, Stanford, California 94305, USA. [15]Department of BioSciences, Rice University, Houston, Texas 77005, USA.

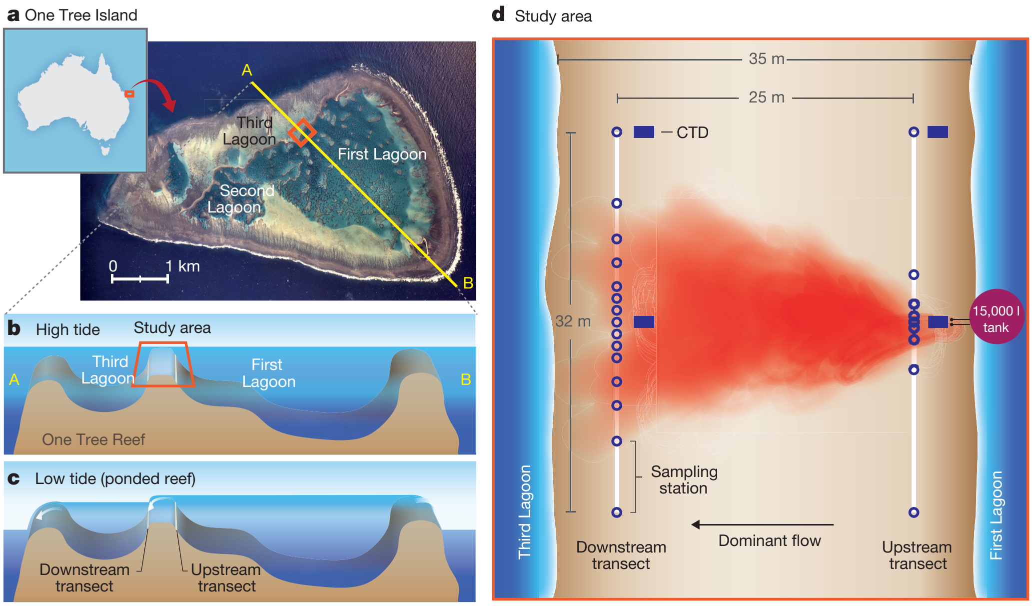


Figure 1 | Study site and experimental design. a, Map of Australia and aerial photograph of One Tree Reef with the study area denoted by an orange square. The map, sourced under Creative Commons CC0, is freely available for commercial use. Use of the photograph was permitted under an Educational license from the University of Sydney. **b**, **c**, Cross-sections of the reef along the yellow line are shown for high (**b**) and low (**c**) tides, demonstrating the unidirectional flow from the upper lagoon (First Lagoon), over the reef flat study area, and into the lower lagoon (Third Lagoon) during low tide. **d**, Schematic of the study area (to scale) indicating the positioning of the transects and sampling locations (blue circles).

and dissolved oxygen concentrations are provided in Extended Data Table 2.

Plots of the alkalinity and dye concentrations along the upstream and downstream transects illustrate the spatial distribution of the plume within the study area (Fig. 3a–d). On control days, when dye but no alkalinity was added, these parameters were not correlated, and the mean alkalinity–dye slopes did not differ from zero (Fig. 3e). On these days, the difference in alkalinity between the upstream and downstream transects was due to background reef calcification and is represented by the difference in y intercepts. On experiment days, when alkalinity and dye were jointly introduced to the study site, these parameters were well correlated, resulting in positive, significantly non-zero alkalinity–dye slopes both for the upstream and for the downstream transects (Fig. 3f). On these days, background reef calcification is represented by the difference in y intercepts (same as control days), and the fraction of added alkalinity taken up by the reef flat, f_{uptake}, was calculated as the difference between the upstream and downstream alkalinity–dye slopes:

$$f_{\mathrm{uptake}} = 1 - (r_{\mathrm{down}}/r_{\mathrm{up}}) \quad (1)$$

where r_{up} and r_{down} are the ratios (slopes) of alkalinity to dye for the upstream and downstream transects, respectively, in μmol kg^{-1} ppb^{-1} or mmol g^{-1}. At a fixed rate of alkalinity and dye addition, r_{up} indicates the amount of added alkalinity entering our study site, while r_{down} indicates the amount of added alkalinity leaving our study site. The difference in these two values indicates the amount of added alkalinity taken up by the reef community and was used to calculate the percentage increase in net calcification according to equations (2)–(4).

Data from all days were analysed using a multivariate regression approach to calculate alkalinity–dye ratios (slopes) and mean background alkalinities (y intercepts) of the upstream and downstream transects, while simultaneously accounting for natural spatial and temporal variability (see Supplementary Information and Extended Data Figs 5–7). Mean alkalinity–dye slopes are presented in Fig. 4a. Results of a mixed-effects model indicate that upstream and downstream slopes are significantly different on experiment days but not control days, rejecting the null hypothesis that net community calcification did not respond to alkalinization (see Supplementary Information).

The fractional uptake of added alkalinity was calculated according to equation (1) and averaged for all control and experimental days. Using this method, we estimate that an average of 17.3% ± 2.3% (1 s.e.m.) of the experimentally added alkalinity was taken up by the reef community.

The percentage increase in net calcification, ΔG, resulting from alkalinity addition was calculated as:

$$\Delta G = G_{\mathrm{increase}}/G_{\mathrm{background}} \quad (2)$$

where G_{increase} is the additional calcification resulting from alkalinity addition in mmol s^{-1}, and $G_{\mathrm{background}}$ is the background calcification in mmol s^{-1} (that is, the calcification rate without added alkalinity). G_{increase} and $G_{\mathrm{background}}$ were calculated as

$$G_{\mathrm{increase}} = P_{\mathrm{dye}}(r_{\mathrm{up}} - r_{\mathrm{down}}) \quad (3)$$

$$G_{\mathrm{background}} = F(Alk_{\mathrm{up}} - Alk_{\mathrm{down}}) \quad (4)$$

where P_{dye} is the pumping rate of the dye in g s^{-1}, F is the volumetric flow rate in m^3 s^{-1}, and Alk_{up} and Alk_{down} are the mean background alkalinities (that is, the y intercepts) of the upstream and downstream transects, respectively, in mmol m^{-3} (see Supplementary Information). Using these equations, we estimate net community calcification increased by an average of 6.9% ± 0.9% (Fig. 4b). A one-tailed, unpaired t-test indicates that the change in calcification on experiment days was significantly greater than control days ($t_{20} = 1.981$, $P < 0.05$). On the basis of laboratory and mesocosm studies[15], the mean response of coral calcification to a unit change in Ω_{arag} is approximately 15%. Throughout the entire study area (inside and outside the plume), Ω_{arag} was elevated by an average of 0.4 units, indicating a theoretical increase in coral calcification of 6%, which agrees closely with the observed increase of 6.9%. Caution must be applied, however, when comparing calcification relationships derived from coral studies[15] to mixed-reef communities such as that of our study site.

The hypothesis that Ω_{arag} exerts strong control over coral reef calcification is supported by laboratory experiments and models[16,17] (but see ref. 18); however, isolating this control in a natural setting is complicated by the multiple drivers of calcification, which are often highly

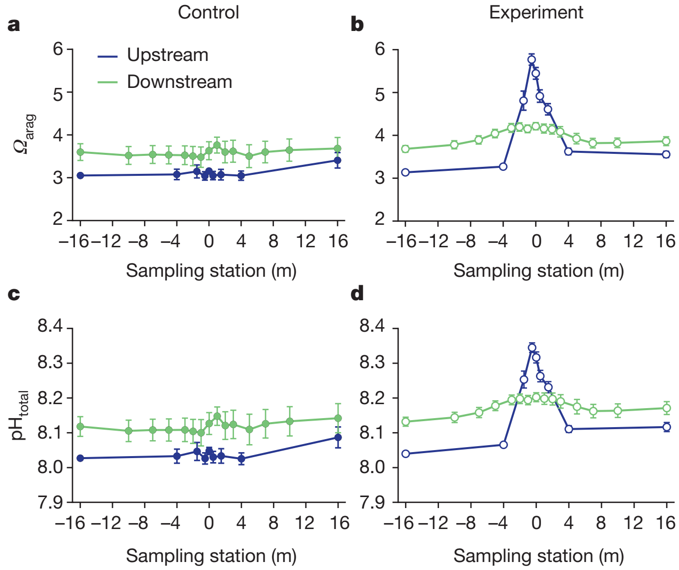


Figure 2 | Chemical conditions for control ($N=7$) and experiment ($N=15$) days (mean ± 1 s.e.m.). **a**, **b**, Aragonite saturation states (Ω_{arag}); **c**, **d**, pH. Error bars are indicative of day-to-day and hour-to-hour variability (not measurement error); estimated measurement errors are smaller than line thickness and are provided in the Methods. Experimental day standard errors are smaller than control day standard errors primarily because of the larger *N*.

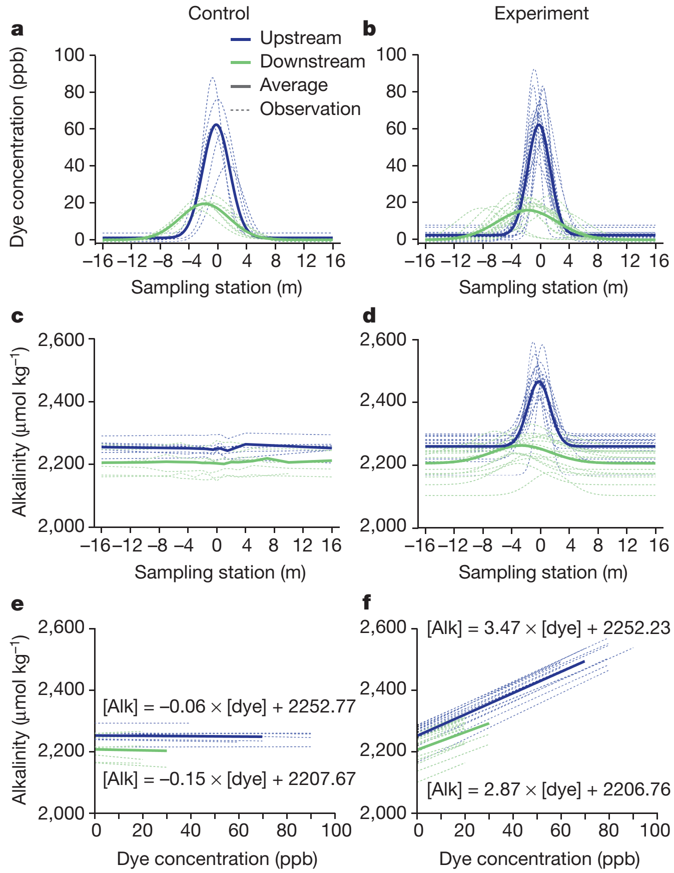


Figure 3 | Relationships between alkalinity and dye for control ($N=7$) and experiment ($N=15$) days. **a**, **b**, Dye concentrations; **c**, **d**, alkalinities; **e**, **f**, alkalinity–dye slopes. **e**, On control days, when dye, but no NaOH, was added to the study site, these parameters are not correlated, and the resulting alkalinity–dye slopes are not significantly different from zero. **f**, On experimental days, dye and NaOH were jointly added to the study site, and the correlation between these parameters results in a positive, significantly non-zero slope. Mean alkalinity–dye slopes for control and experiment days are shown in Fig. 4a.

correlated (for example, production, Ω_{arag}, light, temperature)[6,12,19–24]. Previous attempts to manipulate seawater chemistry in the natural environment were unable to demonstrate a causal relationship between seawater chemistry and reef calcification[25]. Further, retrospective studies documenting declines in coral reef calcification over the past several decades were unable to isolate the influences of various causal factors (for example, ocean warming, acidification, water quality, fishing pressure) owing to the confounding influence of co-varying parameters and a lack of reliable long-term carbonate chemistry observations[7,26]. Our experimental approach demonstrates the influence of alkalinity (and Ω_{arag}) on net community calcification in a natural setting by uncoupling Ω_{arag} from otherwise co-varying confounding environmental factors (where 'uncouple' is used in the technical sense of 'lack of correlation'). We demonstrate that restoring [CO_3^{2-}] and Ω_{arag} closer to pre-industrial values enhances net community calcification, providing evidence that ocean acidification may have contributed to the documented declines in coral reef calcification[6–12] in the industrial era.

Notably, ocean acidification is one of many stressors acting on coral reef calcification. Simultaneously to decreasing Ω_{arag}, sea surface temperatures have warmed by an estimated 0.4–0.8 °C (varying by region) since the early 1800s (ref. 27) which is posited to have increased calcification rates until a recent 'tipping point'[28]. Identifying the relative contributions of various environmental factors, and how they interact, to the documented declines in coral reef calcification is complex yet essential to understanding how calcification will probably change in the coming decades. Further work, using methods developed here, could examine how coral reef response is affected by a variety of stressors (in isolation and combination) and duration of exposure, and help to assess geographic variability in sensitivity to ongoing ocean acidification.

The Ω_{arag} of the tropical oceans is expected to continue declining from 3.8 to approximately 3.0 by the middle of the century and 2.3 by the end of the century[14]. Deliberate alkalinization has been proposed as a geoengineering technique to offset ocean acidification impacts on coral reefs and other shallow marine ecosystems[29]. Our results indicate that this approach could, in principle, help protect coral reefs from ocean acidification; however, the technical challenges associated with implementation would probably make it infeasible at anything but highly localized scales (for example, protected bays, lagoons). Large-scale and long-term protection of marine ecosystems from the threat of ocean acidification depends on deep and rapid reductions in anthropogenic emissions of carbon dioxide[30].

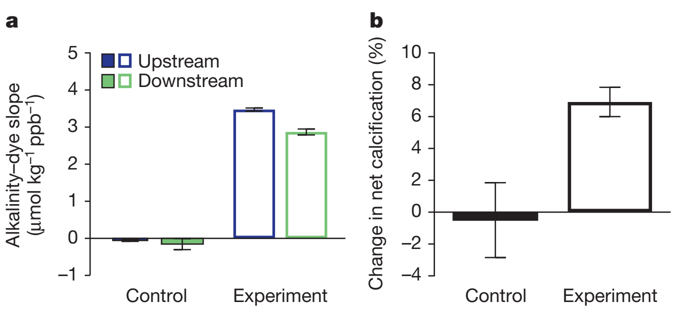


Figure 4 | Alkalinity–dye slopes and percentage change in net calcification for control ($N=7$) and experiment ($N=15$) days (mean ± 1 s.e.m.). **a**, **b**, The difference between upstream and downstream slopes (**a**) was used to calculate the uptake of added alkalinity (equation (1)) and the percentage change in net calcification (**b**) (equations (2)–(4)). The reef community took up an average of 17.3% ± 2.3% of the added alkalinity, implying a 6.9 ± 0.9% increase in net calcification. The percentage change in calcification on experiment days was significantly greater than control days (one-tailed, unpaired *t*-test, $t_{20}=1.981$, $P<0.05$). Results by day are presented in Extended Data Fig. 7.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 September 2015; accepted 20 January 2016.
Published online 24 February 2016.

1. Le Quéré, C. *et al.* Global carbon budget 2014. *Earth System Science Data* **7,** 47–85 (2015).
2. Fabry, V. J., Seibel, B. A., Feely, R. A. & Orr, J. C. Impacts of ocean acidification on marine fauna and ecosystem processes. *ICES J. Mar. Sci.* **65,** 414–432 (2008).
3. Doney, S. C., Fabry, V. J., Feely, R. A. & Kleypas, J. A. Ocean acidification: the other CO_2 problem. *Annu. Rev. Mar. Sci*. **1,** 169–192 (2009).
4. Andersson, A. J. & Gledhill, D. Ocean acidification and coral reefs: effects on breakdown, dissolution, and net ecosystem calcification. *Annu. Rev. Mar. Sci*. **5,** 321–348 (2013).
5. Silverman, J., Lazar, B., Cao, L., Caldeira, K. & Erez, J. Coral reefs may start dissolving when atmospheric CO_2 doubles. *Geophys. Res. Lett.* **36,** (2009).
6. Silverman, J. *et al.* Community calcification in Lizard Island, Great Barrier Reef: a 33 year perspective. *Geochim. Cosmochim. Acta* **144,** 72–81 (2014).
7. De'ath, G., Lough, J. M. & Fabricius, K. E. Declining coral calcification on the Great Barrier Reef. *Science* **323,** 116–119 (2009).
8. Cooper, T. F., De'Ath, G., Fabricius, K. E. & Lough, J. M. Declining coral calcification in massive *Porites* in two nearshore regions of the northern Great Barrier Reef. *Glob. Change Biol.* **14,** 529–538 (2008).
9. Manzello, D. P. Coral growth with thermal stress and ocean acidification: lessons from the eastern tropical Pacific. *Coral Reefs* **29,** 749–758 (2010).
10. Cantin, N. E., Cohen, A. L., Karnauskas, K. B., Tarrant, A. M. & McCorkle, D. C. Ocean warming slows coral growth in the central Red Sea. *Science* **329,** 322–325 (2010).
11. Tanzil, J. T. *et al.* Regional decline in growth rates of massive Porites corals in Southeast Asia. *Glob. Change Biol.* **19,** 3011–3023 (2013).
12. Silverman, J. *et al.* Carbon turnover rates in the One Tree Island reef: a 40-year perspective. *J. Geophys. Res.* **117,** G03023 (2012).
13. Cao, L. & Caldeira, K. Atmospheric CO_2 stabilization and ocean acidification. *Geophys. Res. Lett.* **35,** (2008).
14. Feely, R. A., Doney, S. C. & Cooley, S. R. Ocean acidification: present conditions and future changes in a high-CO_2 world. *Oceanography* **22,** 36–47 (2009).
15. Chan, N. C. & Connolly, S. R. Sensitivity of coral calcification to ocean acidification: a meta-analysis. *Glob. Change Biol.* **19,** 282–290 (2013).
16. Langdon, C. & Atkinson, M. Effect of elevated pCO_2 on photosynthesis and calcification of corals and interactions with seasonal change in temperature/irradiance and nutrient enrichment. *J. Geophys. Res.* **110,** C09S07 (2005).
17. Kroeker, K. J., Kordas, R. L., Crim, R. N. & Singh, G. G. Meta-analysis reveals negative yet variable effects of ocean acidification on marine organisms. *Ecol. Lett.* **13,** 1419–1434 (2010).
18. Cyronak, T., Schulz, K. G. & Jokiel, P. L. The Omega myth: what really drives lower calcification rates in an acidifying ocean. *ICES J. Mar. Sci.* http://dx.doi.org/10.1093/icesjms/fsv075 (2015).
19. Shaw, E. C., Phinn, S. R., Tilbrook, B. & Steven, A. Natural in situ relationships suggest coral reef calcium carbonate production will decline with ocean acidification. *Limnol. Oceanogr.* **60,** 777–788 (2015).
20. Albright, R., Benthuysen, J., Cantin, N., Caldeira, K. & Anthony, K. Coral reef metabolism and carbon chemistry dynamics of a coral reef flat. *Geophys. Res. Lett.* **42,** 3980–3988 (2015).
21. Koweek, D. *et al.* Environmental and ecological controls of coral community metabolism on Palmyra Atoll. *Coral Reefs* **34,** 339–351 (2014).
22. Shaw, E. C., McNeil, B. I. & Tilbrook, B. Impacts of ocean acidification in naturally variable coral reef flat ecosystems. *J. Geophys. Res.* **117,** C03038 (2012).
23. Albright, R., Langdon, C. & Anthony, K. R. N. Dynamics of seawater carbonate chemistry, production, and calcification of a coral reef flat, central Great Barrier Reef. *Biogeosciences* **10,** 6747–6758 (2013).
24. Falter, J. L., Lowe, R. J., Zhang, Z. & McCulloch, M. Physical and biological controls on the carbonate chemistry of coral reef waters: effects of metabolism, wave forcing, sea level, and geomorphology. *PLoS ONE* **8,** e53303 (2013).
25. Kline, D. I. *et al.* A short-term *in situ* CO_2 enrichment experiment on Heron Island (GBR). *Sci. Rep.* **2,** 413 (2012).
26. Helmle, K. P., Dodge, R. E., Swart, P. K., Gledhill, D. K. & Eakin, C. M. Growth rates of Florida corals from 1937 to 1996 and their response to climate change. *Nature Commun.* **2,** 215 (2011).
27. Tierney, J. E. *et al.* Tropical sea surface temperatures for the past four centuries reconstructed from coral archives. *Paleoceanography* **30,** 226–252 (2015).
28. Lough, J. M. & Cooper, T. F. New insights from coral growth band studies in an era of rapid environmental change. *Earth Sci. Rev.* **108,** 170–184 (2011).
29. National Research Council. *Climate Intervention: Carbon Dioxide Removal and Reliable Sequestration* (National Academies Press, 2015).
30. Ricke, K. L., Orr, J. C., Schneider, K. & Caldeira, K. Risks to coral reefs from ocean carbonate chemistry changes in recent earth system model projections. *Environ. Res. Lett.* **8,** 034003 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Dunbar for the use of his laboratory and D. Mucciarone for laboratory training and assistance; the Australian Institute of Marine Science for scientific and technical support; Y. Estrada for graphics assistance; and the following people for their support in the field and/or laboratory: M. Byrne, A. Chai, R. Graham, T. Hill, D. Kline, B. Kravitz, J. Reiffel, D. Ross, E. Shaw, and the staff of the One Tree Island Research Station. Expedition and staff support was provided by the Carnegie Institution for Science. Some additional support for staff, but not expedition expenses, was provided by the Fund for Innovative Climate and Energy Research. This work was permitted by the Great Barrier Reef Marine Park Authority under permit G14/36863.1.

Author Contributions R.A., J.K.M., K.Sc., J.S., and K.C. conceived and designed the project. J.K.M., K.Sc., J.S., J.P., K.L.R., and K.Sh. conducted pilot studies and collected preliminary data. R.A., L.K., L.C., B.M.M., Y.N., T.R., M.S., K.W., A.N., J.H., and K.C. performed the experiments. R.A. and K.C. performed the computational analyses. K.Z. assisted with statistical analyses. R.A. wrote the manuscript with input from K.C. All co-authors reviewed and approved the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.A. (ralbright@carnegiescience.edu).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Concept. The dual tracer regression method developed here is an extension of ref. 31 and may have applications in other areas of research, such as nutrient or pollution assessments, uptake of industrial or agricultural waste, etc. The primary experimental criteria are that the active and passive tracers are added in a fixed ratio and at a fixed rate. The methods described here apply to situations where there is a dominant flow direction, dispersion or dilution, and a need to measure the effect of a reagent on community flux.

Experimental setup. Before low tide each day, a 15 m^3 floating 'header' tank was partly submerged in First Lagoon, adjacent to the reef flat study site. The tank was gravity-fed with ambient seawater from the lagoon, and when necessary, a submersible pump was used to completely fill the tank. Two marine grade bilge pumps (3,000 gallons per hour, Five Oceans) were secured inside the tank for mixing during chemical addition and to deliver the solution to the study site. On 22 days, 4 g Rhodamine WT (20 g of a 20% solution, Turner Designs 10–108), dissolved in 0.5 l reverse osmosis (RO) water, was manually added to the tank over the course of ~30 min and mixed. On 15 of those days, 600 g (~15 mol) of NaOH, dissolved in 1.5 l RO water, was also introduced into the tank. The solution inside the tank was subsequently mixed for an additional 30–45 min to ensure homogeneity. When a strong source of alkalinity is added to seawater, brucite forms as a solid precipitate. However, for pH levels below ~9, brucite dissolves. On the basis of visual inspection and associated laboratory experiments, we estimated brucite dissolution to occur on the timescale of ~100 s. Therefore, a mixing time of 30–45 min was sufficient to ensure complete dissolution of brucite; during this time, a handheld pH probe (Oakton) was used to manually check that the pH of the tank solution did not exceed 9.0. The tank was covered to avoid equilibration with the atmosphere, but given that the tank was emptied over a period of 60 min, it is possible that air–sea fluxes were not completely avoided. While acid/base manipulations of seawater carbonate chemistry are not directly equivalent to the addition/removal of dissolved inorganic carbon (C_T), the differences in carbon speciation between acid/base manipulations and CO_2 gas manipulations are minor if seawater is not allowed to equilibrate with the atmosphere (that is, in closed or continuous-flow systems). Further, it is infeasible to remove CO_2 from large volumes of seawater (>10,000 l) under field conditions. Therefore, acid/base manipulations are considered justified techniques to alter seawater carbonate chemistry in circumstances where large volumes of seawater are being manipulated, such as mimicking natural flow on coral reefs, particularly if the system is not allowed to equilibrate with the atmosphere[32].

The seawater solution from the tank (control days: seawater + dye; experiment days: seawater + NaOH + dye) was pumped onto the reef flat for a period of 60 min starting at the predicted time of low tide (Extended Data Table 1) at a constant rate (~2 l s^{-1}). The solution was introduced to the study site via the two bilge pumps that were submerged in the tank, connected to two lengths of (1.5-inch inner diameter) vinyl tubing that were secured to a cinder block located ~2 m upstream from the centre of the study site. Throughout the addition, the ratio of alkalinity to dye being added to the study site was assumed to be constant, and flow within the study site was considered to be in steady-state. On 30 September 2014 (a control day), dye was added using a peristaltic pump instead of the above configuration; this was because low tide occurred at 5:32, and assembly of the tank configuration was not possible in low-light conditions.

Following a pumping period of 60 min, discrete water samples were taken at defined sampling locations along the length of two parallel transects that defined the borders of the study area: one along the upstream edge (adjacent to First Lagoon) and the other along the downstream edge (adjacent to Third Lagoon, see Extended Data Fig. 3). The total width of the reef flat in this area is approximately 35 m, and the upstream and downstream transects were separated by ~25 m. The length of each transect was 32 m. Thus, the study area consisted of an approximate 25 m × 32 m rectangle (800 m^2). The upstream transect consisted of 9 sampling stations spanning the width of 32 m, and the downstream transect consisted of 15 sampling stations spanning the width of 32 m. Sample locations were strategically assigned with a higher density near the centre of the study area to accurately characterize the shape of the resulting alkalinity and dye plume/curve. Spacing of the station locations is depicted in Extended Data Fig. 3.

Following the 60 min pumping period, discrete samples were collected at each of the 24 sample locations by pumping reef water into 500 ml borosilicate glass bottles (Corning, 1500-500 Pyrex glass reagent bottle) using battery-operated liquid transfer pumps (Sierra Tools, model JB5684). To minimize chemical variation due to minor changes in sampling depth and/or location, precise sample locations were marked with plastic discs, nailed to the reef substrate. Samples were collected along the upstream and downstream transects simultaneously by five individuals, with each person sampling four or five locations. All samples were typically collected in less than 3 minutes, and it was assumed that the study site was in steady state during this time (that is, all fluxes and flows did not change during the 3-minute sampling interval). Samples were immediately returned to the One Tree Island Research Station, where they were subsampled and analysed for pH, total alkalinity (A_T), dissolved inorganic carbon (C_T), and rhodamine (see 'Chemical Analyses' section). For three upstream stations (−U16, U0, U16) and three downstream stations (−D16, D0, D16), nitrate and ammonia concentrations were also determined. See Extended Data Fig. 3 for station locations.

CTD (conductivity–temperature–depth) devices (YSI models 6600, 6920) were deployed at four sampling locations, two upstream (−U16, U0) and two downstream (−D16, D0) for continuous measurements of seawater temperature, salinity, depth, and dissolved oxygen concentration. These instruments logged continuously at 2-minute intervals over the 22 study days. Discrete water samples (Corning, 1500-250 Pyrex glass reagent bottle) were collected each day at each of the four CTD locations, and the dissolved oxygen concentration was measured using an automated potentiometric Winkler titration technique[33]. These values were used to verify CTD measurements.

Alkalinity–dye slopes, r, and mean background alkalinities, $\hat{a}$, for each day were calculated using paired alkalinity and dye measurements that were collected across all sampling stations, transects, and days (see Supplementary Information). Over a 4-week period, we conducted our experimental protocol 23 times: 8 control days and 15 experimental days. One control day was omitted from subsequent analyses owing to intense rain that heavily influenced alkalinity measurements, resulting in 7 control days and 22 total days. This resulted in a total of 526 paired alkalinity and dye measurements that were used in the fitting procedure described in the Supplementary Information. Two previous expeditions to One Tree Island (September/October 2012 and March 2013) characterized site variability and allowed testing of the methods presented here. Preliminary data generated in these expeditions indicated that demonstrating statistical significance was dependent on maximizing signal (uptake of experimentally added alkalinity) to noise (natural/background uptake of alkalinity).

On experiment days, the difference between the upstream and downstream alkalinity–dye slopes indicates the fraction of experimentally added alkalinity that was taken up by the reef (equation (1) of the main text). We analysed the difference between slopes using a mixed-effects model in R (see Supplementary Information). Comparison of confidence intervals indicates that upstream and downstream slopes are significantly different on experiment days but not on control days. Shapiro–Wilk W-tests were used to verify the underlying assumptions of normality ($P > 0.05$). The purpose of control days was to demonstrate that significant changes in alkalinity–dye slopes do not occur when NaOH is not added, and to characterize natural spatial and temporal variability in the study site. Further, with this study methodology, effectively, within experimental days, the part of the study site that is not affected by the alkalinity-rich plume serves as additional control for the part of the study site that is affected by the plume.

While reef processes other than calcification can alter seawater alkalinity (for example, changes in nutrients, salinity), a previous study showed that changes in salinity and nutrients had a negligible effect on changes in alkalinity in coral reefs[34]. Salinity and nutrient data from our study are provided in Extended Data Table 2.

Code availability. The Mathematica routine used to calculate alkalinity-to-dye ratios (slopes) and dye-free mean alkalinity estimates (y intercepts) for each day is provided in the Supplementary Information.

Chemical analyses. Discrete samples were immediately returned to the laboratory on One Tree Island where they were analysed for pH_{total}, total alkalinity (A_T), and rhodamine, and subsampled for the later determination of total dissolved inorganic carbon (C_T), and nutrients (NH_4^+, NO_2, and NO_3). All measurements and calculations were consistent with 'best practices' recommendations[35]. For 99.6% of station–day combinations (24 stations × 22 days = 528 bottles), we successfully measured pH_{total}, A_T, rhodamine, and C_T, resulting in 526 paired measurements.

Aragonite saturation state (Ω_{arag}), carbonate ion concentration ($[CO_3^{2-}]$), and p_{CO_2} were calculated as a function of A_T, pH_{total}, and *in situ* salinity and temperature using the program CO2SYS[36]; dissociation constants for carbonate and boric acid were determined as in ref. 37 and as refitted in ref. 38, and the dissociation constant for boric acid was determined as in ref. 39.

Parameters that were measured at a subset of sampling stations (that is, temperature, salinity, and dissolved oxygen measured at -U16, U0, -D16, D0; nutrients measured at –U16, U0, U16, -D16, D0, D16) are presented in Extended Data Table 2. Parameters that were measured (or calculated) across all sampling stations are presented in Fig. 2 (Ω_{arag} and pH) and Extended Data Fig. 4 (CO_3^{2-}, p_{CO_2}, and C_T). All chemistry data are included in Supplementary Table 1.

Total alkalinity, A_T. Samples for A_T were pre-filtered using GF/F filters (Whatman) and analysed in triplicate using a Metrohm 855 Robotic Titrosampler (Metrohm

USA) using certified 0.1 N HCl (Fisher Chemical) diluted to a nominal concentration of 0.0125 N. Acid was calibrated by analysing certified reference material (CRM, batch 138) from A. Dickson's laboratory before each titration session (twice daily). A_T by volume (μmol l^{-1}) was converted to A_T by mass (μmol kg^{-1}) by applying a density correction using *in situ* salinities and temperatures. For each set of triplicate analyses, data points that were $>10\,\mu$mol kg^{-1} away from the median were removed from the analysis; these outliers resulted from drop scale variability in sample delivery. The resulting mean and standard error were calculated for each sample location on each day. Instrumental precision from 55 analyses of CRM (batch 138) over the course of the study was $<2\,\mu$mol kg^{-1} (1 s.d.). Alkalinities were normalized to the mean salinity of 35.75; salinity-normalized alkalinities were used for subsequent analyses.

pH. Seawater pH_{total} was determined using an Ocean Optics spectrophotometer with 10 cm path length optical cells and *m*-cresol purple dye (Sigma Aldrich), following the methods of ref. 41. Water samples were kept in a temperature-controlled water bath (Thermo Scientific, Precision Microprocessor Controlled 280 Series) at 25 °C before analysis to minimize temperature-induced errors in absorbance measurements. The temperature of each sample was recorded immediately after analysis using a digital thermometer accurate to $\pm$0.05 °C (VWR, Traceable Platinum Ultra-Accurate Digital Thermometer). CO2SYS[36] was used to calculate *in situ* pH values using *in situ* salinity and temperature measurements. Average precision from triplicate measurements for this system was less than 0.010 units (1 s.d.). CRM analyses (TRIS buffer, batch 22, A. Dickson) revealed that the system was accurate to within 0.005 pH units.

Rhodamine WT. Rhodamine WT concentration was measured fluorometrically using a Turner 10AU fluorometer and 25 ml cuvettes. A series of eight standards was made by mass-diluting a 400 ppb (Parts per 10^9) Rhodamine WT standard (Turner Designs) to 0, 0.5, 1, 2, 4, 16, 32, and 64 ppb. The standard curve was measured at the beginning, middle, and end of each measuring day to check for drift. Water samples were kept in a temperature-controlled water bath (Thermo Scientific, Precision Microprocessor Controlled 280 Series) at 25 °C before analysis to minimize temperature-induced errors in fluorescence. The temperature of each sample was recorded immediately after analysis using a digital thermometer accurate to $\pm$0.05 °C (VWR, Traceable Platinum Ultra-Accurate Digital Thermometer). Rhodamine concentrations were temperature-corrected using the formula $F_r = F_s\exp(k(T_s - T_r))$, where F_r and F_s are the fluorescence at the reference and sample temperatures, T_r and T_s, and $k = 0.026$/K, equating to a 2.6% correction per kelvin (ref. 40). Temperature corrections were applied before normalizing values to the standard curve. Dye concentrations were then normalized to the mean salinity of 35.75; salinity-normalized concentrations were used for subsequent analyses. Instrumental precision from triplicate measurements for this system was less than 0.1 ppb.

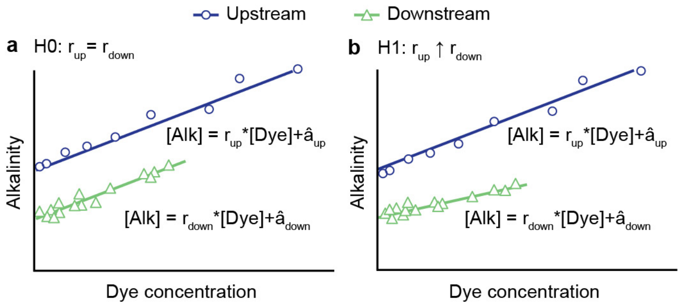
Dissolved inorganic carbon (C_T). C_T samples were subsampled into 30 ml glass serum bottles (Wheaton, 223743), poisoned with 15 μl saturated $HgCl_2$ (0.05% by volume to inhibit biological activity[41]), sealed with rubber stoppers, crimped closed with aluminium caps, and transported to Stanford University for analysis. Samples were analysed approximately 3 months after sampling. C_T was extracted from samples by acidifying them with phosphoric acid (H_3PO_4, 5%) using a custom-built, automated acidification and delivery system (D. Mucciarone) using high-grade nitrogen as a carrier gas connected to an infrared gas analyser (Licor 7000). All samples were analysed in duplicate. The instrument was calibrated daily using CRM (Batches 141, 138), provided by A. Dickson. Immediate duplicate analyses of samples usually yielded instrumental precision of 1–2 μmol kg^{-1}.

Nutrients. Nutrient samples were subsampled into 15 ml conical centrifuge tubes (Falcon). Ammonia samples were immediately frozen, and total ammonium concentrations ($NH_{3\text{-tot}} = NH_3 + NH_4^{1+}$) were later determined using a modified fluorometric method[42]. Nitrate samples were preserved with 0.1 ml 1 N HCl, closed, shaken, and left in the dark at room temperature ($\sim$22 °C) until transport to Eilat, Israel. Nitrite (NO_2^{1-}) was measured using a colorimetric method[43], with a Flow Injection Autoanalyzer (Lachat Instruments model QuickChem 8500). Nitrate (NO_3^{1-}) was measured by reducing it to nitrite using a copperized cadmium column. Precision of ammonia, nitrite, and nitrate measurements was $\sim$ 0.05 μmol l^{-1}. Nitrite and nitrate in this study are reported as total oxidized nitrogen ($TON = NO_2^{1-} + NO_3^{1-}$). Results are provided in Extended Data Table 2.

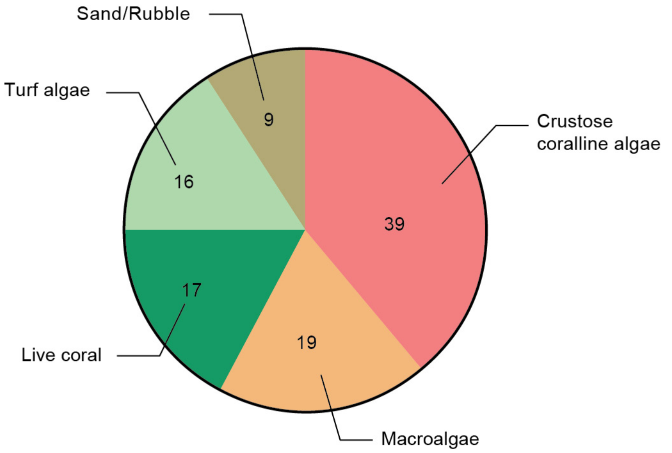
Salinity. Following the first 5 days of observations, it became evident that conductivity measurements from three of the four CTDs proved unreliable; we believed this to be from the formation of oxygen bubbles on the sensors (resulting from high productivity on the reef flat). Therefore, starting on day six, discrete water samples were taken each day at each of the four CTD locations. Samples were stored in an air-conditioned, shaded room until transport to the Australian Institute of Marine Science for analysis on a Guildline Portasal Salinometer (model 8410A), with a precision of $\pm$0.0001 units. Accuracy was verified using CRM (OSIL, IAPSO Standard Seawater, batch P155). For days without discrete salinity measurements ($N = 5$), salinity values were calculated for the upstream transect by developing a linear relationship between the salinometer values and the reliable CTD. Salinities for the downstream transect were calculated from upstream values by applying an offset of 0.08 parts per thousand: this offset represents the mean increase in salinity between the upstream and downstream transects as a result of evaporation (Extended Data Table 2).

Benthic community structure. Benthic surveys were conducted to characterize the underlying community structure of the study area. Five 25 m transects were laid on the reef flat perpendicular to the reef front, spaced approximately 8 m apart. Photographs were taken of 0.25 m^2 quadrats at 5 m intervals. Photographs were analysed with Coral Point Count software with Excel extensions (CPCe) using 25 random points per quadrat. The benthos was assigned to one of six categories: (1) live coral; (2) macroalgae; (3) turf algae; (4) *Halimeda*; (5) crustose coralline algae (CCA); and (6) sand/rubble. Where morphological forms of $CaCO_3$ (for example, rubble, $CaCO_3$ rock) were covered with biologically active groups (for example, turf, CCA), the biologically active group was scored. Results are provided in Extended Data Fig. 2.

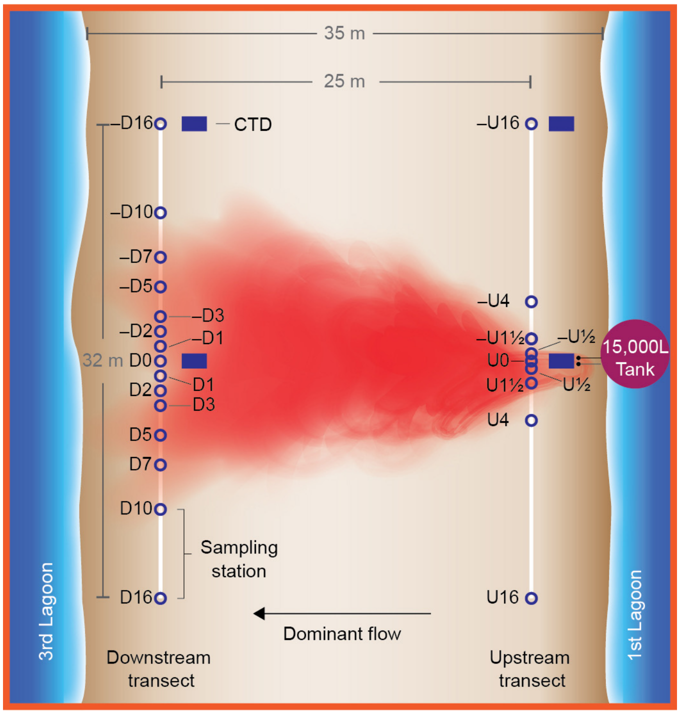
31. Friedlander, S. K., Turner, J. R. & Hering, S. V. A new method for estimating dry deposition velocities for atmospheric aerosols. *J. Aerosol Sci.* **17,** 240–244 (1986).
32. Andersson, A. J. & Mackenzie, F. T. Revisiting four scientific debates in ocean acidification research. *Biogeosciences* **9,** 893–905 (2012).
33. Anderson, L. G., Haraldsson, C. & Lindegren, R. Gran linearization of potentiometric Winkler titrations. *Mar. Chem.* **37,** 179–190 (1992).
34. Kinsey, D. Alkalinity changes and coral reef calcification. *Limnol. Oceanogr.* **23,** 989–991 (1978).
35. Riebesell, U., Fabry, V. J., Hansson, L. & Gattuso, J. P. *Guide to Best Practices for Ocean Acidification Research and Data Reporting* (Publications Office of the European Union, 2010).
36. Lewis, E. & Wallace, D. W. R. Program developed for CO2 system calculations (US Department of Energy Oak Ridge National Laboratory, 1998).
37. Mehrbach, C., Culberson, C. H., Hawley, J. E. & Pytkowicz, R. M. Measurement of the apparent dissociation constants of carbonic acid in seawater at atmospheric pressure. *Limnol. Oceanogr.* **18,** 897–907 (1973).
38. Dickson, A. G. & Millero, F. J. A comparison of the equilibrium constants for the dissociation of carbonic acid in seawater media. *Deep-Sea Res.* **34,** 1733–1743 (1987).
39. Dickson, A. G. Thermodynamics of the dissociation of boric acid in synthetic seawater from 273.15 to 318.15 K. *Deep-Sea Res.* **37,** 755–766 (1990).
40. Wilson, J. F. in *Techniques for Water Resources Investigations of the U.S. Geological Survey,* Book 3, Ch. A12 (U.S. Government Printing Office, 1968).
41. Dickson, A. G., Sabine, C. L. & Christian, J. R. *Guide to Best Practices for Ocean CO_2 Measurements* (North Pacific Marine Science Organization, 2007).
42. Holmes, R. M., Aminot, A., Kerouel, R., Hooker, B. A. & Peterson, B. J. A simple and precise method for measuring ammonium in marine and freshwater ecosystems. *Can. J. Fish. Aquat. Sci.* **56,** 1801–1808 (1999).
43. Grasshoff, K., Kremling, K. & Ehrhardt, M. (eds) *Methods of Seawater Analysis* (Wiley-VCH, 1999).



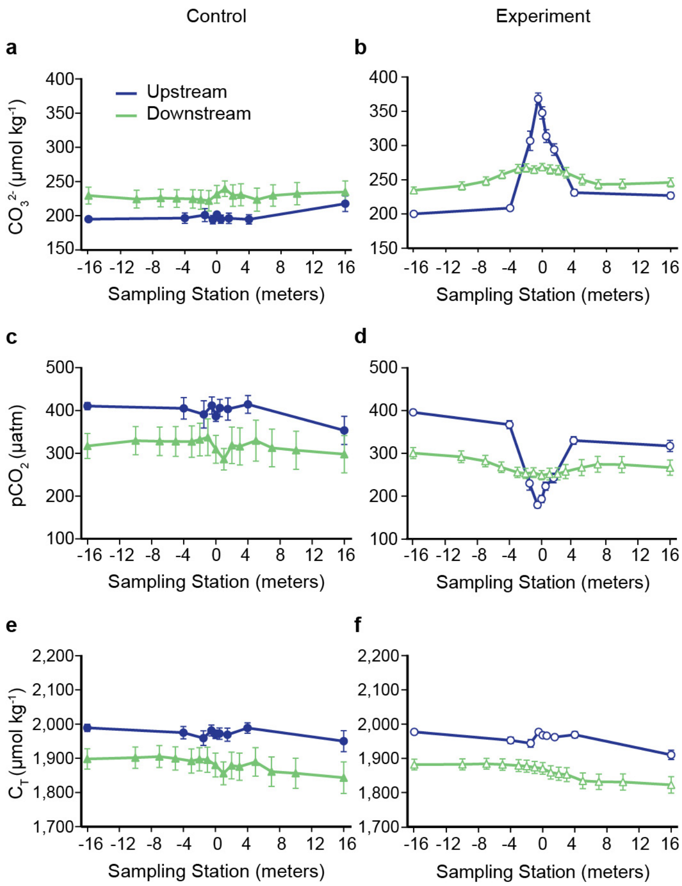
Extended Data Figure 1 | Theoretical representations of the null, H0, and alternative, H1, hypotheses. **a**, In H0, the reef does not take up added alkalinity; here, the change in alkalinity between the upstream and downstream transects would not be systematically related to the dye concentration, and the ratio of the alkalinity–dye relationship, *r*, would not be expected to change between the upstream and downstream locations (that is, $r_{up} = r_{down}$). **b**, In H1, reef uptake of added alkalinity occurs; here, areas with more alkalinity (and more dye) change at a different rate than areas with less alkalinity (and less dye), resulting in a change in the alkalinity–dye slope (that is, $r_{up} > r_{down}$).



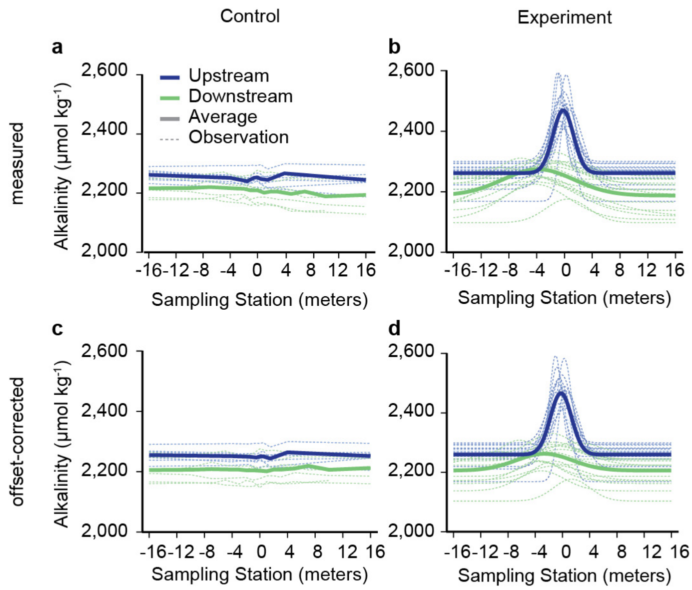
Extended Data Figure 2 | Community composition of the reef flat study area. Percentage cover by benthic type is as follows: crustose coralline algae (39%), live coral (17%), turf algae (16%), macroalgae (19%), sand/rubble (9%), and *Halimeda* (5%).



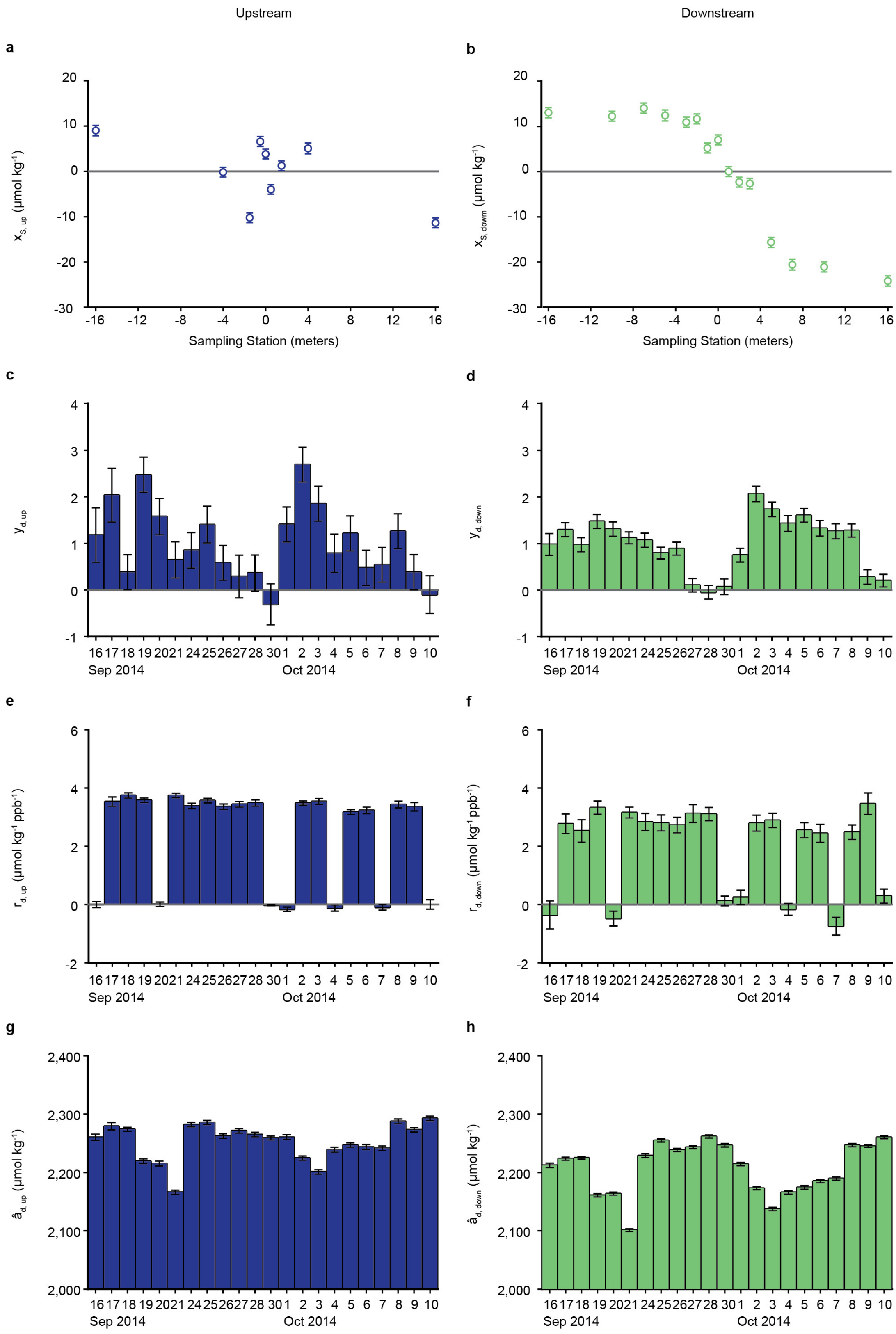
Extended Data Figure 3 | Schematic of study area showing meter-spacing of station locations for the 9 upstream (U) stations and 15 downstream (D) transects. Numbers indicate the metre-spacing from the centre of the study area, denoted as U0 for the upstream transect and D0 for the downstream transect. The outermost sampling locations for the upstream (−U16, U16) and downstream (−D16, D16) transects define the four outermost corners of the study area and were strategically positioned to lie outside the alkalinity–dye plume, rendering zero dye concentrations and added alkalinity.



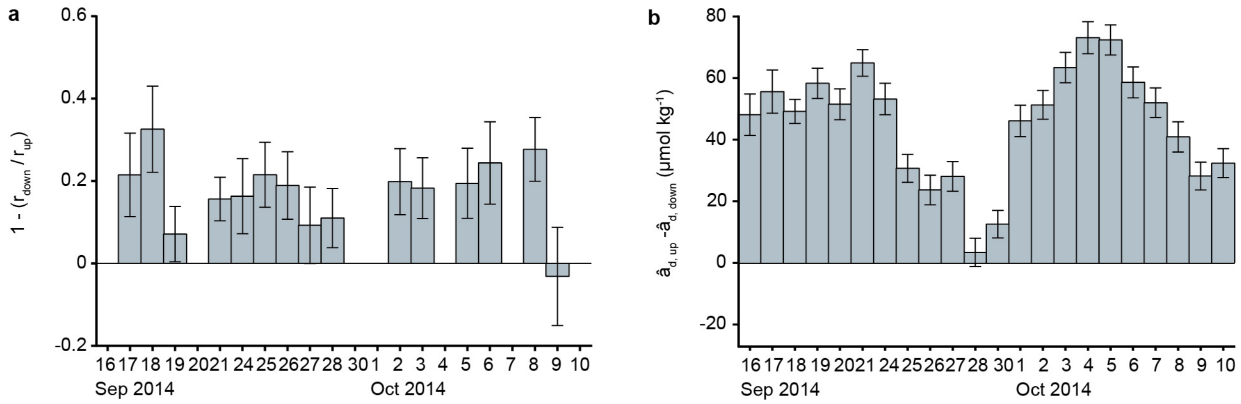
Extended Data Figure 4 | Mean chemical conditions for control ($N = 7$) and experiment ($N = 15$) days. a, b, Carbonate ion concentrations ([CO_3^{2-}]); **c, d,** p_{CO_2}; **e, f,** dissolved inorganic carbon concentrations (C_T) for upstream and downstream transects. Error bars, which represent standard errors, are indicative of day-to-day and hour-to-hour variability (not measurement error); estimates of measurement error are provided in the Methods. Total alkalinity (A_T), dye concentration, aragonite saturation state (Ω_{arag}), and total pH (pH_T) are provided in Figs 2 and 3.



Extended Data Figure 5 | Comparison of alkalinity values before and after 'offset-corrections' used in the multivariate regression analysis. **a**, **b**, Measured (that is, 'raw') alkalinity values. **c**, **d**, 'Offset-corrected' alkalinity values. Bold lines represent average conditions; dashed lines show results by day. See Supplementary Information.



Extended Data Figure 6 | Results of the multivariate regression analysis. **a**, **b**, Unique offsets by station, x_s, for the upstream and downstream transects. **c**, **d**, Magnitude of offsets by day, y_d, for upstream and downstream transects. **e**, **f**, Alkalinity–dye ratios by day, r_d, for upstream and downstream transects. **g**, **h**, Mean background alkalinities by day, $\hat{a}_d$, for upstream and downstream transects. Error bars represent standard errors. See Supplementary Information.



Extended Data Figure 7 | Results of the multivariate regression were used to calculate the additional alkalinity uptake (that is, G_{increase}) and background alkalinity uptake (that is, $G_{\mathrm{background}}$) by day. a, Fraction of added alkalinity taken up by the reef by day, given by $(1 - (r_{\mathrm{down}}/r_{\mathrm{up}})$, equation (1) of main text). **b**, Background reef uptake by day, given by $(\hat{a}_{d,\,\mathrm{up}} - \hat{a}_{d,\,\mathrm{down}})$. Error bars represent standard errors. See Supplementary Information.

Extended Data Table 1 | Schedule for control and experiment days, including date, time, predicted height of low tide, and mean photosynthetically active radiation (PAR) for the 1 h study period

Date	Low Tide (HHMM)	Water Depth (m)	PAR (μmol m^{-2} s^{-1})	Type of Study
16 Sept	0737	0.88	1146	Control
17 Sept	0907	0.93	1626	Experiment
18 Sept	1026	0.85	1797	Experiment
19 Sept	1125	0.72	1800	Experiment
20 Sept	1210	0.59	1715	Control
21 Sept	1249	0.47	1215	Experiment
24 Sept	1425	0.29	931	Experiment
25 Sept	1455	0.30	675	Experiment
26 Sept	1528	0.35	551	Experiment
27 Sept	1604	0.44	297	Experiment
28 Sept	1647	0.58	25	Experiment
30 Sept	0532	0.57	277	Control
01 Oct	0639	0.75	866	Control
02 Oct	0809	0.84	1459	Experiment
03 Oct	0940	0.80	1055	Experiment
04 Oct	1051	0.65	1916	Control
05 Oct	1150	0.46	1636	Experiment
06 Oct	1240	0.30	1707	Experiment
07 Oct	1327	0.19	1453	Control
08 Oct	1411	0.15	1179	Experiment
09 Oct	1455	0.19	787	Experiment
10 Oct	1539	0.29	513	Control

Tides were provided courtesy of One Tree Research Station. The low-tide time represents the time at which pumping onto the reef started; sampling occurred 60 min afterwards. Data for photosynthetically active radiation were obtained from the Australian Institute of Marine Science weather station at One Tree Island (http://data.aims.gov.au/aimsrtds/station.xhtml?station=131). There was no significant difference between mean light levels for control (1,166 ± 217 μmol m^{-2} s^{-1}, mean ± s.e.m.) and experiment (1,098 ± 150 μmol m^{-2} s^{-1}) days.

Extended Data Table 2 | Mean (±1 s.e.m.) values for temperature (*T*), salinity (*S*), ammonium (NH_4), nitrite and nitrate ($NO_2 + NO_3$), and dissolved oxygen (DO) during the 22-day study period

	Control		Experiment	
	Upstream	Downstream	Upstream	Downstream
T (^oC)	23.0 ± 0.3	23.5 ± 0.5	23.3 ± 0.2	23.7 ± 0.3
S	35.79 ± 0.02	35.84 ± 0.02	35.71 ± 0.02	35.80 ± 0.02
NH_4 (μmol L^{-1})	0.40 ± 0.02	0.23 ± 0.02	0.37 ± 0.01	0.24 ± 0.01
$NO_2 + NO_3$ (μmol L^{-1})	1.14 ± 0.04	0.80 ± 0.02	1.08 ± 0.02	0.78 ± 0.01
DO (mg L^{-1})	5.9 ± 0.2	7.1 ± 0.4	6.2 ± 0.1	7.3 ± 0.1

Note that underlying natural variability (that is, day-to-day, hour-to-hour) contributes to standard errors; measurement errors for each parameter are indicated in the Methods.

doi:10.1038/nature16960

Vast assembly of vocal marine mammals from diverse species on fish spawning ground

Delin Wang[1], Heriberto Garcia[1], Wei Huang[1], Duong D. Tran[1], Ankita D. Jain[2], Dong Hoon Yi[2], Zheng Gong[1,2], J. Michael Jech[3], Olav Rune Godø[4], Nicholas C. Makris[2] & Purnima Ratilal[1]

Observing marine mammal (MM) populations continuously in time and space over the immense ocean areas they inhabit is challenging but essential for gathering an unambiguous record of their distribution, as well as understanding their behaviour and interaction with prey species[1–6]. Here we use passive ocean acoustic waveguide remote sensing (POAWRS)[7,8] in an important North Atlantic feeding ground[9,10] to instantaneously detect, localize and classify MM vocalizations from diverse species over an approximately 100,000 km^2 region. More than eight species of vocal MMs are found to spatially converge on fish spawning areas containing massive densely populated herring shoals at night-time[11–16] and diffuse herring distributions during daytime. We find the vocal MMs divide the enormous fish prey field into species-specific foraging areas with varying degrees of spatial overlap, maintained for at least two weeks of the herring spawning period. The recorded vocalization rates are diel (24 h)-dependent for all MM species, with some significantly more vocal at night and others more vocal during the day. The four key baleen whale species of the region: fin, humpback, blue and minke have vocalization rate trends that are highly correlated to trends in fish shoaling density and to each other over the diel cycle. These results reveal the temporospatial dynamics of combined multi-species MM foraging activities in the vicinity of an extensive fish prey field that forms a massive ecological hotspot, and would be unattainable with conventional methodologies. Understanding MM behaviour and distributions is essential for management of marine ecosystems and for accessing anthropogenic impacts on these protected marine species[1–5,17,18].

Here we present the ecosystem-wide spatial distributions of vocal MMs from multiple cetacean species, including both mysticetes and odontocetes, acquired simultaneously with that of their fish prey during their feeding season in the Gulf of Maine (GOM), an important North Atlantic MM foraging ground[9,10] (Figs 1 and 2). The GOM is among the more diverse, productive, and trophically complex marine temperate areas in the world, with Atlantic herring (*Clupea harengus*) comprising a keystone prey species, common in the diets of many marine mammals, piscivorous fish and seabirds of the region[9]. Using a large-aperture, densely sampled, coherent hydrophone array with orders of magnitude higher array gain[19] than previously available in acoustic marine mammal sensing, we could detect, localize and classify vocalizing MMs from multiple species instantaneously over an approximately 100,000 km^2 region by POAWRS[7,8] without aliasing[19] in time and space (Fig. 3a, Methods and Extended Data Figs 1–5). Simultaneous fish distributions were acquired instantaneously over tens of thousands of square kilometre areas by ocean acoustic waveguide imaging (OAWRS)[11–14], combined with conventional fisheries ultrasonic echosounding[15] and fish trawl sampling[16] to obtain thousands of calibrations at statistically significant locations[13]. We mapped the

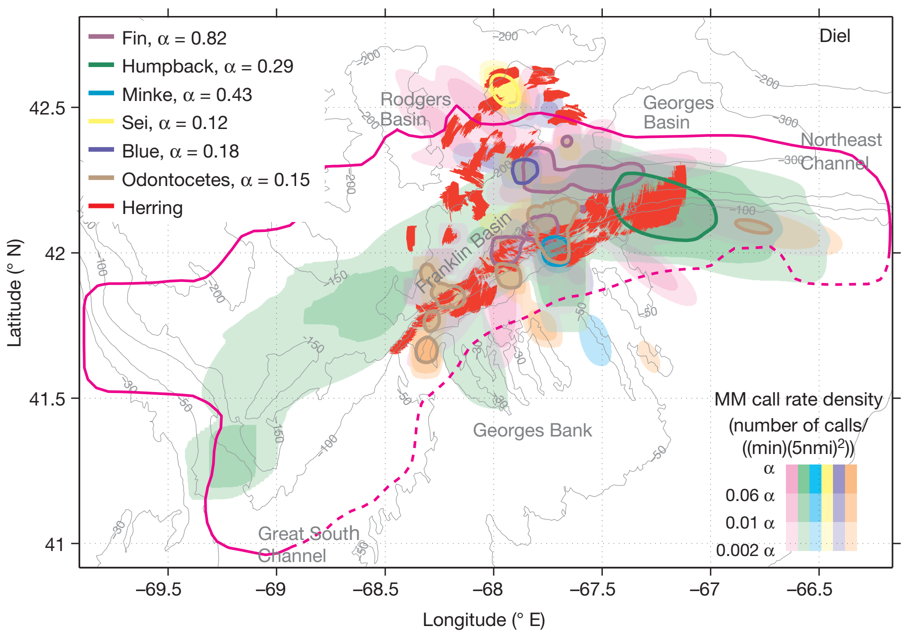


Figure 1 | Full diel cycle distributions of MM vocalizations and fish. Vocalizing MMs from diverse species are convergent on spawning herring distributions in autumn 2006 (26 September to 6 October). Dense Atlantic herring shoals (>0.2 fish per m^2, red shaded areas) imaged using OAWRS system[12,13] and diffuse herring populations (approximately 0.053 fish per m^2, bounded by magenta line) obtained from conventional fish finding sonar[15,16]. The MM call rate densities in units of number of calls per minute per 25 nmi^2 ((min)(5 nmi)^2) measured by POAWRS have peak values α indicated. Detailed shoaling herring and MM species vocalization spatial distributions can be found in Extended Data Fig. 6 and Supplementary Information section III, respectively. The bathymetric data (contours shown in grey) are obtained from the US National Centers for Environmental Information.

[1]Laboratory for Ocean Acoustics and Ecosystem Sensing, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115, USA. [2]Laboratory for Undersea Remote Sensing, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [3]Northeast Fisheries Science Center, 166 Water Street, Woods Hole, Massachusetts 02543, USA. [4]Institute of Marine Research, Post Office Box 1870, Nordnes, N-5817 Bergen, Norway.

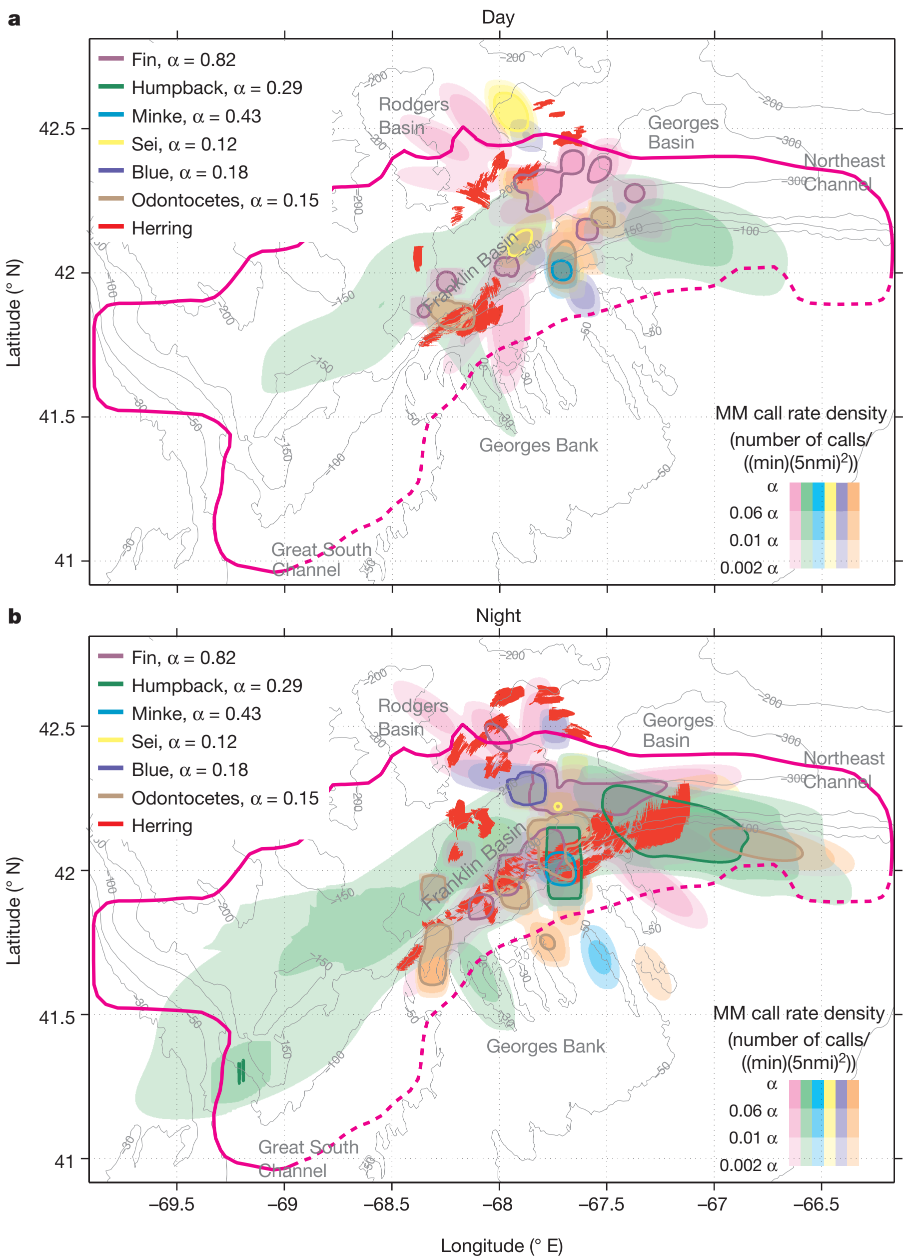


Figure 2 | Day and night distributions of MM vocalizations and fish. **a**, **b**, Daytime (**a**) and night-time (**b**) distributions of vocalizing MMs from diverse species. Regions with dense herring populations (>0.2 fish per m^2, red shaded areas) during day and night are shown. The daytime hours are between sunrise and sunset (06:00 to 18:00 EDT), while night-time hours are between sunset and sunrise the next day (18:00 to 06:00 EDT). All other contours are similar to Fig. 1. The bathymetric data (contours shown in grey) are obtained from the US National Centers for Environmental Information.

vocalization temporospatial distributions of five baleen whale species (Mysticeti): blue (*Balaenoptera musculus*), fin (*Balaenoptera physalus*), humpback (*Megaptera novaeangliae*), sei (*Balaenoptera borealis*) and minke (*Balaenoptera acutorostrata*), as well as toothed whale species (Odontoceti): sperm (*Physeter macrocephalus*), pilot (*Globicephala* spp.), orca (*Orcinus orca*) and several other delphinid species. We examined whether these vocal MMs had species-specific spatial affinities and how they varied across the prey field over the diel cycle. We further investigated the extent to which MM vocalization behaviour for each species is correlated to fish shoaling density which had recently been shown to be a major factor driving humpback vocal behaviour in their feeding ground[7].

The overall vocalization rate spatial distributions of the five baleen whale species and the toothed whales are found to be highly focused on dense herring shoaling areas[12–15] on the northern flank of Georges Bank, and almost entirely (more than 90% of the calls) contained within the region of at least diffuse herring aggregations[15,16] that extends over a roughly 12,000 km^2 area between Georges and Rodgers Basins in the north, Georges Bank in the south, Northeast Channel to the east, and Great South Channel to the west (Figs 1 and 2). The dense herring shoals[12–14] are characterized by 0.2 fish per m^2 at shoal boundaries to over 10 fish per m^2 at shoal centres, factors 4 to 200 times greater than those of the diffuse aggregations[12–15] that are characterized by roughly 0.053 fish per m^2 (Methods, Extended Data Fig. 6). The cumulative call-rate distribution for all MM species fall off rapidly with increasing range from the dense herring shoals (Fig. 3b). A significant majority of MM vocalizations from any species, at least 63% corresponding to the *e*-folding decay value of the cumulative call-rate distribution in range, originate in areas that either completely overlap with or lie within 3–8 km range

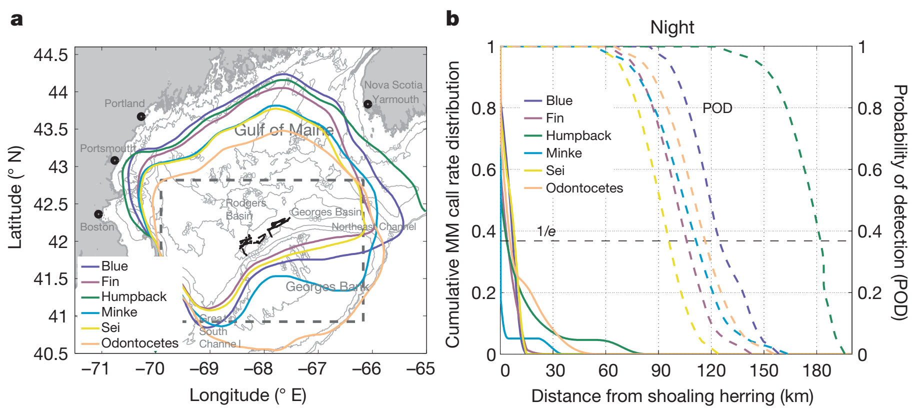


Figure 3 | POAWRS MM detection region and cumulative nocturnal MM call rate distribution. **a**, POAWRS 0.5 probability of detection (POD) contour for MM vocalizations in the Gulf of Maine. Dotted box is the region shown in Figs 1 and 2. **b**, Cumulative nocturnal MM vocalization rate distribution as a function of minimum distance from nocturnal herring shoaling areas. The *e*-folding distances of the cumulative nocturnal MM vocalization rate distributions range from 0–8 km depending on the species. In contrast, the azimuthally-averaged POAWRS MM POD *e*-folding distances from nocturnal herring shoaling areas are a factor of 10 to 100 times larger. The bathymetric data (contours shown in grey) are obtained from the US National Centers for Environmental Information.

of dense herring shoals. These *e*-folding distances are less than the size of the herring shoals and can be traversed by most MMs within several minutes to at most an hour, timescales small relative to the roughly 12 h duration of the herring shoal's nocturnal existence. In contrast, the zooplankton distribution in the area of intense MM vocalizations is diffuse, with volumetric densities (715 ± 550 zooplankton per m^3, Supplementary Information section V) roughly nine times smaller than those found in areas where baleen whales have been previously observed to be actively feeding[20,21] on zooplankton ($6{,}500 \pm 2{,}000$ zooplankton per m^3). From an energetics perspective, the dense herring populations, with biomass ranging from 5 g m^{-3} to over 250 g m^{-3}, provide orders of magnitude more efficient prey than the diffuse zooplankton of approximately 0.6 ± 0.5 g m^{-3} biomass. The spatial relationship quantified here between regions of dense MM vocalization rate densities and those of dense fish shoaling densities are consistent with a mechanism by which MMs from around the GOM converge on localized dense prey fields that are uniquely available during the annual herring spawning season for efficient feeding.

We find the vocal MMs are not uniformly distributed over the prey field, but concentrated in species-specific population centres with varying degrees of spatial overlap that can vary over the diel cycle depending on the species (Figs 1 and 2 and Supplementary Information section III). The dominant MM consumers of GOM herring have been proposed[9] to be fins and humpbacks, accounting for close to 50% of the estimated 150 ± 50 kilotons of herring consumed annually by MM predators, with the remaining consumed by other MM species, such as minkes and various delphinids. MMs are responsible for 50% of total herring consumption by non-human predators[9]. The Atlantic herring that spawn on northern Georges Bank comprise roughly 15% to 25% of the GOM Georges Bank complex herring stock[16]. These images reveal the previously uncharted spatial affinities of multiple MM species simultaneously engaged in foraging activities in the vicinity of an enormous fish prey field. The ability to observe predator distributions with respect to changing prey fields is essential for advancing our understanding of ecosystem processes[20,22,23].

The vocalization repertoire of MMs is highly species dependent, varying in acoustic frequency content, duration, and repetition rate (see Methods for MM species vocal classification). Of the mysticete whales, the fins are found to be the most vocal group with the highest measured vocalization rates of roughly 14,000 calls per day, followed by humpbacks with roughly 2,000 calls per day. The measured vocalization rates for all MM species are provided in Extended Data Table 1. MMs vocalize for a variety of purposes[4,5,8,23–29] that include songs associated with mating in mysticete whales, social sounds associated with inter- and intragroup interactions, signals for long range or night-time acoustic communication, calls for coordinated movement during feeding or migration, and echolocation for prey-finding in odontocetes that may possibly extend to certain mysticete whale species (Methods).

Determining the temporal and spatial variations in MM vocalizations is important for understanding their natural behaviours[4,25] and also crucial for providing controls when assessing anthropogenic effects on that behaviour[5]. All five baleen whale species and the toothed whales analysed here have vocalization rates that vary over the diel cycle. Four of the five baleen whale species, fin, humpback, blue and minke, are found to be more vocal at night (from dusk 18:00 Eastern daylight time (EDT) through to dawn 06:00 EDT the next day) with night-time vocalization rates a factor of between 2 to 10 times that of the daytime (Extended Data Table 1). In contrast, the sei whales and collectively, the odontocetes, are more vocal in the daytime by a factor of roughly 1.25. The vocalization rate spatial distribution is strongly diel-dependent for some MM species, but less so for others (Fig. 2a, b). Combined sensing with POAWRS and OAWRS[7,8,11–14] enables MM and fish distributions to be simultaneously monitored over instantaneous continental-shelf scale regions, and presents a significant advantage in areal coverage over conventional line-transect visual[10,30] and ultrasonic echosounding survey techniques respectively.

We correlated the MM vocalization rate time series with the herring shoaling areal population density time series over the diel cycle as a function of MM species (Fig. 4, Extended Data Fig. 7 and Extended Data Table 1). Four of the five vocal baleen whale species, fin, humpback, minke and blue, are found to have temporal call rate trends that are highly correlated to temporal trends in herring shoaling density and also to each other. For humpbacks, their night-time vocalizations are dominated by downsweep 'meows', 'feeding cries'and 'bow-shaped' calls that are associated with fish-feeding activities, occur ten times more frequently at night than in the daytime accounting for the high correlation ($r_{\mathrm{MM,fish}} = 0.87$) to nocturnal herring shoaling densities[7]. The high temporal correlation ($r_{\mathrm{MM,fish}} = 0.82$) obtained here between fin vocalization rate and herring shoaling density indicates that the factor of 2 increase in the characteristic 20 Hz fin vocalizations at night is likely associated with increased fish-feeding activities. This is supported by the spatial focusing of the fin vocalization distribution on north-central Georges Bank in the location of dense herring shoaling populations during night-time (Fig. 2). The minke vocalization rates are well correlated to herring shoaling densities, both temporally ($r_{\mathrm{MM,fish}} = 0.64$) and spatially. The minke call rate distribution spatial overlap with dense herring shoaling areas increases from 0% in the day to 71% at night (Fig. 3b and Extended Data Fig. 8). This implies

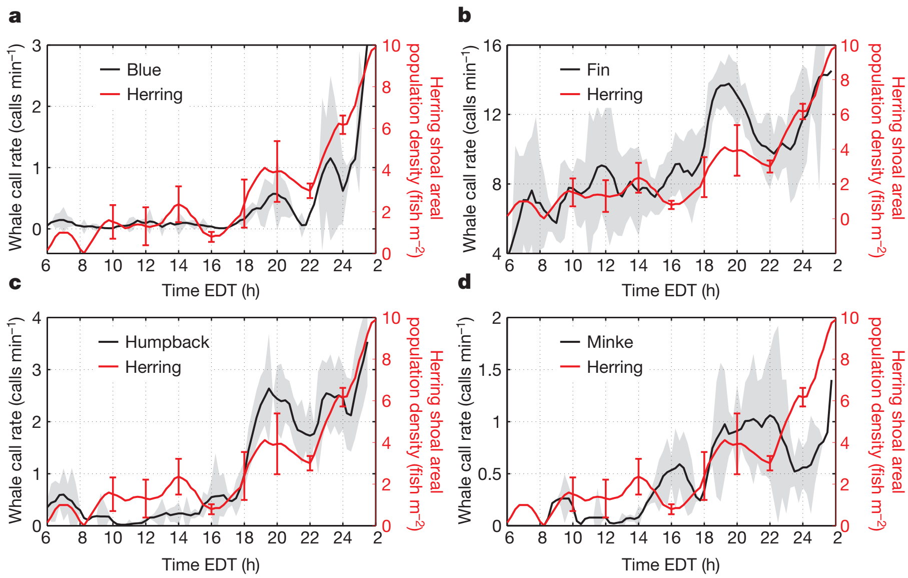


Figure 4 | Diel MM call rate and herring shoal areal population density time series. a–d, Mean call rates for blue (**a**), fin (**b**), humpback (**c**), and minke (**d**) whales are correlated to Atlantic herring shoal mean areal population density over the diel cycle. The error bars indicate standard deviations obtained from averaging the time series over multiple diel cycles from 26 September to 6 October 2006. The period from roughly 2–6 EDT contains a data gap.

that the fivefold increase in minke buzz vocalization sequences recorded here, which resemble the odontocete echolocation click sequences (compare Extended Data Fig. 1d, e), play an important role in minke fish-feeding activity. Here we find the blue vocalization rates have the highest temporal correlation ($r_{\mathrm{MM,fish}} = 0.91$) to herring shoaling densities. The blue whale calls we detected are comprised predominantly of short-duration (approximately 2 s) audible downsweep (type D[26]) calls, which were previously observed for blue whales in both the North Atlantic[26] and Pacific feeding grounds[27] and regarded as contact calls. The high temporal correlation between blue whale call rates and herring densities obtained here suggest that it may be possible that blue whales are consuming herring and the type D vocalizations are feeding-related, though this fish-feeding behaviour has not been observed for blue whales. (Other possible explanations are discussed in Methods.) Previous observations of sei whales in the GOM found lower call rates at night for their downsweep chirp signals attributed to zooplankton-feeding and higher call rates in the daytime while engaged in social interactions and less feeding[23]. Here we find a similar pattern for their vocalization behaviour in the vicinity of dense herring shoals, which may result from more fish-feeding at night than in daytime. The tooth whales collectively have vocalization rates that are temporally uncorrelated to herring shoaling densities, perhaps because their fish-feeding strategies differ from those of baleen whales. The toothed whales hunt small fish groups and feed on individual fish[28,29], whereas baleen whales engulf large quantities, hundreds to thousands of fish, all at once[24] and rely on fish shoaling together for efficient feeding.

The ecosystem-scale predator–prey behaviour documented here is expected to be a regular feature of the annual herring spawning season on Georges Bank and an important aspect of the life cycles of the various species involved that is likely to be found in other ocean ecosystems where many of the same conditions occur.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 June; accepted 21 December 2015.
Published online 2 March 2016.

1. Tittensor, D. P. *et al.* Global patterns and predictors of marine biodiversity across taxa. *Nature* **466,** 1098–1101 (2010).
2. Schofield, O. *et al.* How do polar marine ecosystems respond to rapid climate change? *Science* **328,** 1520–1523 (2010).
3. Block, B. A. *et al.* Tracking apex marine predator movements in a dynamic ocean. *Nature* **475,** 86–90 (2011).
4. Noad, M., Cato, D., Bryden, M., Jenner, M. & Jenner, K. Cultural revolution in whale songs. *Nature* **408,** 537 (2000).
5. Miller, P. J., Biassoni, N., Samuels, A. & Tyack, P. L. Whale songs lengthen in response to sonar. *Nature* **405,** 903 (2000).
6. Kaschner, K., Quick, N. J., Jewell, R., Williams, R. & Harris, C. M. Global coverage of cetacean line-transect surveys: status quo, data gaps and future challenges. *PLoS ONE* **7,** e44075 (2012).
7. Gong, Z. *et al.* Ecosystem scale acoustic sensing reveals humpback whale behavior synchronous with herring spawning processes and re-evaluation finds no effect of sonar on humpback song occurrence in the Gulf of Maine in fall 2006. *PLoS ONE* **9,** e104733 (2014).
8. Tran, D. D. *et al.* Using a coherent hydrophone array for observing sperm whale range, classification, and shallow-water dive profiles. *J. Acoust. Soc. Am.* **135,** 3352–3363 (2014).
9. Overholtz, W. J. & Link, J. S. Consumption impacts by marine mammals, fish, and seabirds on the Gulf of Maine–Georges Bank Atlantic herring (*Clupea harengus*) complex during the years 1977–2002. *ICES J. Mar. Sci.* **64,** 83–96 (2007).
10. Waring, G. T., Josephson, E., Maze-Foley, K. & Rosel, P. E. U.S. Atlantic and Gulf of Mexico Marine Mammal Stock Assessments 2014. *NOAA Tech Memo NMFS NE* Vol. 231 (2014).
11. Makris, N. C. *et al.* Fish population and behavior revealed by instantaneous continental shelf-scale imaging. *Science* **311,** 660–663 (2006).
12. Makris, N. C. *et al.* Critical population density triggers rapid formation of vast oceanic fish shoals. *Science* **323,** 1734–1737 (2009).
13. Gong, Z. *et al.* Low-frequency target strength and abundance of shoaling Atlantic herring (*Clupea harengus*) in the Gulf of Maine during the ocean acoustic waveguide remote sensing 2006 experiment. *J. Acoust. Soc. Am.* **127,** 104–123 (2010).
14. Jagannathan, S. *et al.* Ocean acoustic waveguide remote sensing (OAWRS) of marine ecosystems. *Mar. Ecol. Prog. Ser.* **395,** 137–160 (2009).
15. Jech, J. M. & Stroman, F. Aggregative patterns of pre-spawning Atlantic herring on Georges Bank from 1999–2010. *Aquat. Living Resour.* **25,** 1–14 (2012).
16. 54th Northeast Regional Stock Assessment Workshop (54th SAW) Assessment Report. Part A. Atlantic herring stock assessment for 2012. *US Dept Commer, Northeast Fish Sci. Cent.* Ref. Doc. 12–18; 600 p. http://www.nefsc.noaa.gov/publications/ (2012).
17. deYoung B., Heath, M., Werner, F., Chai, F., Megrey, B. & Monfray, P. Challenges of modeling ocean basin ecosystems. *Science* **304,** 1463–1466 (2004).
18. Overholtz, W. & Link, J. A simulation model to explore the response of the Gulf of Maine food web to large-scale environmental and ecological changes. *Ecol. Modell.* **220,** 2491–2502 (2009).
19. Kay, S. M. Fundamentals of statistical signal processing, Vol. II: Detection Theory. (Prentice Hall, 1998).
20. Mayo, C. A. & Marx, M. K. Surface foraging behavior of the North Atlantic right whale, *Eubalaena glacialis*, and associated zooplankton characteristics. *Can. J. Zool.* **68,** 2214–2220 (1990).

21. Dolphin, W. F. Prey densities and foraging of humpback whales, *Megaptera novaeangliae. Experientia* **43,** 468–471 (1987).
22. Sims, D. W. & Quayle, V. A. Selective foraging behavior of basking sharks on zooplankton in a small scale front. *Nature* **393,** 460–464 (1998).
23. Baumgartner, M. F. & Frantantoni, D. M. Diel periodicity in both sei whale vocalization rates and the vertical migration of their copepod prey observed from ocean gliders. *Limnol. Oceanogr.* **53,** 2197–2209 (2008).
24. Nøttestad, L., Fern, A., Mackinson, S., Pitcher, T. & Misund, O. A. How whales influence herring school dynamics in a cold-front area of the Norwegian Sea. *ICES J. Mar. Sci.* **59,** 393–400 (2002).
25. Janik, V. M. Whistle matching in wild bottlenose dolphins (*Tursiops truncatus*). *Science* **289,** 1355–1357 (2000).
26. Berchok, C. L., Bradley, D. L. & Gabrielson, T. B. St. Lawrence blue whale vocalizations revisited: characterization of calls detected from 1998 to 2001. *J. Acoust. Soc. Am.* **120,** 2340–2354 (2006).
27. Wiggins, S. M., Oleson, E. M., McDonald, M. A. & Hildebrand, J. A. Blue whale (*Blaenoptera musculus*) diel call patterns offshore of Southern California. *Aquat. Mamm.* **31,** 161–168 (2005).
28. Similä, T. Sonar observations of killer whales *(Orcinus orca)* feeding on herring schools. *Aquat. Mamm.* **23,** 119–126 (1997).
29. Wiirsig, B. Delphinid foraging strategies in *Dolphin cognition and Behavior: a Comparative Approach* 347–359 (Psychology Press, 1986).
30. Buckland, S. T. *et al. Introduction to Distance Sampling: Estimating Abundance of Biological Population* (Oxford Univ. Press, 2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements Permission for this National Oceanographic Partnership Program experiment was given in the Office of Naval Research document 5090 Ser 321RF/096/06. This research was supported by the US National Science Foundation, the US Office of Naval Research (Ocean Acoustics Program), the National Oceanographic Partnership Program, the US Presidential Early Career Award for Scientists and Engineers, the Alfred P. Sloan Foundation, the Census of Marine Life, and Northeastern University. The authors thank J. R. Preston for assistance with GOM 2006 experiment, D. H. Cato and P. L. Tyack for discussions.

Author Contributions Overall concept and approach conceived and developed by P.R. Implementation, data analysis and interpretation directed by P.R., conducted by D.W., W.H., H.G. and D.D.T. with contributions from A.D.J., D.H.Y. and Z.G. The GOM 2006 experiment data collection was led by N.C.M., P.R. and J.M.J. The article was written by P.R. with contributions from D.W., W.H., J.M.J., O.R.G. and N.C.M. All authors read and discussed the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.R. (purnima@ece.neu.edu).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

POAWRS receiver array recordings of MM vocalizations. Acoustic recordings were acquired using a 160 hydrophone-element horizontal receiver line-array[31] towed by a research vessel along designated tracks north of Georges Bank from 19 September to 6 October 2006 (GOM 2006 experiment)[11–14], coinciding with the annual herring spawning period[15,16,32] on the northern flank of Georges Bank[9,33,34]. Here we use data from all 160 hydrophone elements nested[13,31] into four subapertures, where each subaperture contains 64 hydrophones for spatially and temporally unaliased sensing up to 4 kHz (sampling rate of POAWRS was 8 kHz)[35]. Acoustic pressure-time series measured by sensors across the receiver array were converted to two-dimensional (2D) beam-time series by time-domain beamforming[19,36], and further converted to spectrograms by temporal Fourier transform. MM vocalizations were automatically extracted using a threshold detector (>5.6 dB signal-to-noise ratio (SNR)) with a detection rate of roughly 87% ± 5% depending on the species, call characteristics and in-beam ambient noise levels. The beamformed spectrograms were subsequently checked by visual inspection to improve expected detection accuracy to over 92%. Examples of typical vocalization spectrogram from diverse MM species are shown in Extended Data Fig. 1.

The high gain[19,36] of the densely-sampled large aperture coherent POAWRS receiver array used here, up to $10 \log_{10} n = 18$ dB gain where $n = 64$ hydrophones for each sub-aperture, enabled detection of whale vocalizations either two orders of magnitude more distant in range or lower in SNR than a single hydrophone which has no array gain (Extended Data Fig. 2). The angular resolution of the receiver array[12] is dependent on the measured bearing, array aperture length and acoustic wavelength, tabulated in Table 1 of ref. 13 for select frequencies.

Characteristics and function of MM vocalizations from diverse species. The time-frequency characteristics of each MM vocalization was extracted via pitch tracking[37–39] and applied to classify the vocalization according to species (Extended Data Fig. 3 and Extended Data Table 2). A combination of extracted features, orthogonalized via principle component analysis (PCA), were used to optimize the vocalization species classification employing *k*-means and Bayesian-based Gaussian mixture model clustering approaches[40]. The bearing-time trajectories of each closely associated series of vocalizations were also taken into account to ensure consistent classification, and to minimize the automatic classification error to between 0.5% to 7% depending on the species.

In the low frequency range from 10 Hz to 100 Hz, the acoustic spectra were dominated by fin, blue and sei vocalizations (Extended Data Fig. 1a–c). The fins were identified from their characteristic 20 Hz centre frequency calls[41–44] that have been associated with communication among fin individuals[45] and also found to be uttered by males as breeding displays in their mating grounds[41,46]. Given the large volume of fin vocalizations measured here in the vicinity of dense shoaling fish populations, averaging 14,000 calls per day, these 20 Hz calls can also be associated with feeding behaviour, serving as communication signals or for coordination among individuals in their foraging ground. The blues were identified from their audible downsweeps — type D calls, burps and grunts[26,47–50], previously found to be vocalized by both sexes, regarded as contact or social calls[48] produced by individuals at shallow depths[47] of 10–40 m. The seis were identified from their downsweep calls occurring singly or as doublets with roughly 4 s separation, and also sometimes as triplets[23,51,52], hypothesized to be long-range contact calls potentially enabling coordinated activities such as feeding or breeding[52].

The spectra in the mid frequency range from 100 Hz to 1,000 Hz were dominated by minke and humpback vocalizations. The minkes were identified from their buzzes comprised of a series of high and low frequency click sequences[53–55], has characteristics similar to the highly repetitive pulse train of odontocetes that may be suitable for prey echolocation. Compared to other baleen whales, the humpbacks have a fairly extensive vocalization repertoire. Here the humpbacks were identified from their songs[4,56], as well as non-song calls[57,58] with characteristics provided in ref. 7. Male humpbacks vocalize songs which are patterned sequences of calls as breeding displays[4] in their mating ground, and have been observed to carry the tunes into their feeding grounds[59–61] (Extended Data Fig. 2). The non-song vocalizations detected include 'feeding cries' similar to those observed in Alaskan humpback cooperative group feeding on herring schools[57], as well as 'bow-shaped' calls and 'meows' suited for night time communication[7] among humpback individuals and coordination during group feeding activities.

The spectra at frequencies higher than 1 kHz were dominated by odontocete vocalizations (Extended Data Fig. 1e–h). They consist of sperm whale slow and usual click, and creak sequences[8,62,63], pilot and killer whale whistles, as well as a wide range of repetitive sequences of downsweep chirp signals roughly 0.7 s duration with varying bandwidths between 200 to 1,000 Hz, all occurring above 1 kHz that can be attributed to pilot[64,65] or killer whales[66–68], or a variety of other delphinid species[25]. The highly repetitive click sequences used for prey echolocation occur at frequencies higher than 10 kHz for many odontocete species, beyond our sampling frequency range. The largest of them, the sperm whale, has slow and usual click and creak sequences with significant energy as low as 1 kHz[8,69]. The whistles and wide variety of downsweep chirp signals we recorded in the frequency range of odontocete vocalizations may serve as contact calls between individuals and to facilitate cohesion during foraging or travel[64,65,70]. The fin, humpback, sei, minke, pilot and orca whales, and common and bottlenose dolphin species were visually sighted during the experiment.

Determination of diel MM vocalization rate time series. The bearing-time trajectories of vocalizations from multiple MM species received by the POAWRS receiver array are shown for two days in Extended Data Fig. 4. For humpbacks, the alternation from song to non-song calls in their vocalization repertoire over several diel cycles are plotted as a function of bearing-time trajectory in Extended Data Fig. 5 for comparison. The line-array's left–right ambiguity is resolved following the approach outlined in refs. 7, 71. The bearing ranges from 100° to 240° from true north for our array spans Georges Bank from east to west respectively. The diel vocalization rate (calls per min) time series shown in Fig. 4 and Extended Data Fig. 7 for each MM species is obtained by averaging the daily vocalization rate time series for that species over the entire experiment. The MM vocalization rate time series, initially calculated in 15 min bins, are averaged over a 1.25 h running window corresponding to the half power width quantifying the temporal correlation scale of the fish shoaling density time series. For both minke and odontocete whale click sequences, since the duration of each click sequence is highly variable from a few seconds to over a minute, the call rates shown in Fig. 4 and Extended Data Fig. 7 represent the number of 5 s intervals that contain click sequences.

Localization and call rate spatial distributions of diverse vocalizing MM species. The horizontal location of each MM vocalization consists of a range and a bearing estimate. The moving array triangulation (MAT)[71] and the array invariant (AI)[71–73] methods were applied to determine the range of the vocalizations from the horizontal receiver array centre. Position estimation error, or the root mean squared (RMS) distance between the actual and estimated location, is a combination of range and bearing errors quantified for this array in refs. 7, 71 and 73. Range estimation error, expressed as the percentage of the range from the source location to the horizontal receiver array centre, for the MAT technique is roughly 2% at array broadside and gradually increases to 10% at 65° from broadside and 25% at 90° from broadside, that is, near or at endfire[71]. Range estimation error for the AI method is roughly 4% to 8% over all azimuthal directions[71,73]. Bearing estimation error of the time domain beamformer is roughly 0.5° at broadside and gradually increases to 6.0° at endfire[71]. These errors are determined at the same experimental site and time period as the MM position estimates presented here, from thousands of controlled source signals transmitted by the same source array used to locate the herring shoals presented here and are based on absolute global positioning system (GPS) ground truth measurements of the source array's position[71], which are accurate to within 3 m to 10 m. More than 80% of vocalizing MMs are found to be located between 0° to 65° from the broadside direction of the horizontal receiver array. Position estimation error[7,71] is less than 2 km for majority of the vocalizing MMs localized in Figs 1 and 2 since they are found within roughly 50 km of the horizontal receiver array centre. This error is over an order of magnitude smaller than the spatial scales of the MM concentrations shown in Figs 1 and 2, and consequently has negligible influence on the analyses and results.

The estimated locations for all MM calls over the duration of our data collection are used to generate the call rate density map for each MM species shown in Figs 1 and 2 following the approach described in ref. 7. The location of each call is characterized by a 2D Gaussian probability density function with mean equal to the measured mean position from MAT or the AI method and standard deviations determined by the measured range and bearing standard deviations. The MM call rate density map for each species is determined by superposition of the 2D spatial probability densities for the location of each call, normalized by the total measurement time. (See Supplementary Information section III for more detailed diel, diurnal and nocturnal MM spatial vocalization rate distributions).

Estimating MM detection region for POAWRS receiver array. The probability of detection $P_D(r)$ of the MM vocalizations from each species as a function of range r from the POAWRS receiver array, shown in Fig. 3 and Extended Data Fig. 8, are calculated employing the formulation[7,13,74–86] and parameters[7,87–89] provided in Supplementary Information section I. Higher transmission loss occurs in shallower waters due to more intense and pervasive bottom interaction[86,90–92] and mode stripping effects, especially at the low frequency vocalization range of large baleen whales. Transmission loss in deeper waters is typically significantly lower due to upward refraction[86,90] which leads to far less intense and pervasive bottom

interaction, as is the case in the deeper waters north of Georges Bank[86,90–92]. As noted in ref. 7, highly directional transmission loss may then occur when there are large depth variations about a receiver. This effect makes the detection range of whales in directions to the north of our receiver and Georges Bank typically much greater than in directions to its south where the relatively shallow waters of Georges Bank are found. The fact that we localized the sources of many whale calls at great distances along shallow water propagation paths on Georges Bank in directions where transmission loss was greater and found negligibly small vocalization rates in the deeper waters north of Georges Bank where transmission loss was much less (Fig. 3a), greatly emphasizes the finding that the vocalization rates originating from the region further north of Georges Bank were negligibly small.

The dominant portion of the whale population within the POAWRS detection region is expected to occur in areas with dense vocalization rates for each MM species. This is because the probability of detecting no vocalizations in a region where a MM species is abundant over our two week observation period is negligibly small. This implies that there are insignificant numbers of MMs within the detection region at locations with low call rates or no calls. This is consistent with acoustic-based marine mammal population density estimation, which makes the standard assumption that there are no whales expected to be present in a time-space region where there are no calls[93–98], or that this number is negligibly small. The MM vocalization rate spatial distribution obtained here showing the northern flank of Georges Bank and its immediate vicinity as an autumn season MM hotspot is consistent with analysis of three decades (1970–2005) of visual line transect survey data for the entire GOM region[10,99] (Supplementary Information section IV).

Atlantic herring areal population density distribution and time series. The Atlantic herring instantaneous areal population density over wide areas shown in Figs 1 and 2 and Extended Data Fig. 6 were obtained from active OAWRS imaging after extensive calibration with tens of thousand instantaneous coincident conventional ultrasonic fisheries echosounding measurements[12–14], with fish species identification and physiological parameters extracted from trawl samples collected over the course of the experiment[15]. The Atlantic herring shoals were consistently observed to form when the population density reached a critical value of 0.2 fish per m^2 where correlated behaviour began, following simple physical theories[12]. This critical density was also consistently found to be the boundary where the dense shoal ended and diffuse populations that did not engage in correlated behaviour began[12–14]. Populations within the dense shoals were variable from 0.2 to over 10 fish per m^2. These shoals extended roughly 20–60 m vertically in water-column depths of 80–200 m. In contrast, the diffuse fish populations approximately 0.053 fish per m^2 were found close to (within 3–5 m of) the seafloor. The shoals formed near sunset and persisted until near sunrise, starting on the northern flank of Georges Bank and migrating southward to shallower waters on the bank. This diurnal behavioural pattern was consistently observed during our roughly two week measurement time period[12–14].

The annual autumn season Atlantic herring spawning activity on the northern flank of Georges Bank has been recorded by the US National Marine Fisheries Services (NMFS) for over 30 years, coinciding their survey of the Georges Bank herring stock with this period each year[9,15,16,32], including our GOM 2006 experiment. The overall Georges Bank Atlantic herring stock estimate for autumn 2006 based on OAWRS[16] survey has been found to match well (within 10% to 20%) with independent NMFS stock estimates for 2006 and 2007 (ref. 16).

Midwater trawl hauls were conducted on an ad hoc basis to sample significant backscatter observed during the NMFS acoustic survey in autumn 2006 (ref. 15). Stomach-content volume and diet composition of up to 15 herring per trawl haul (length-stratified subsampling) were recorded at sea. The majority of pre-spawning herring were not feeding as evidenced by the high proportion of empty stomachs[15]. During the acoustic/midwater trawl survey on Georges Bank in 2006, 95% of herring had nothing in their stomach (that is, stomach-content volume was 0 cm^3), which is comparable to other years where the proportion of empty stomachs ranged from 100% in 2003 to 69% in 2010.

The areal resolution of the fish density distribution shown in Figs 1 and 2 obtained from the OAWRS imaging system is 30 m in range after matched filtering and averaging, and varies between 150 m to roughly 1.5 km in cross-range for the vast majority of fish hotspots included here, where the cross-range resolution is dependent on the array angular resolution, bearing, and fish range. The MM call rate distributions have spatial resolutions bounded by roughly 2 km that is of the same order of magnitude as the fish areal population density distribution in cross range.

We combine the herring shoaling activity on Georges Bank, where massive and highly dense shoals occur predominantly during the night on the northern flank[7,8,11–14], with that for the region to the north of the Bank between Rodgers and Georges Basins, where we observe less persistent herring groups forming throughout the diel cycle[15]. The areal fish population density (in fish/m^2) time series shown Fig. 4 and Extended Data Fig. 7 is obtained by averaging the daily herring shoaling areal population density time series initially calculated in 15 min bins, then averaged over a 1.25 h running window corresponding to the half power width quantifying the temporal correlation scale of the fish shoaling density time series.

The temporal correlation coefficients $r_{\mathrm{MM,fish}}$ in Extended Data Table 1 quantify the degree of similarity between the diel MM call rate time series $c_{\mathrm{MM}}(t)$ and the diel fish areal population density time series $n_{\mathrm{fish}}(t)$ and are calculated[19,100] via

$$r_{MM,fish}=\frac{\sum_{k=1}^{N}(c_{MM}(t_k)-\overline{c_{MM}})(n_{fish}(t_k)-\overline{n_{fish}})}{\sqrt{\sum_{k=1}^{N}(c_{MM}(t_k)-\overline{c_{MM}})^2}\sqrt{\sum_{k=1}^{N}(n_{fish}(t_k)-\overline{n_{fish}})^2}} \qquad (1)$$

using measurements $c_{\mathrm{MM}}(t_j)$ and $c_{\mathrm{MM}}(t_k)$ that are independent for $j \neq k$, so that $t_{k+1}-t_k=1.25$ h. Similarly, the temporal correlation coefficient $r_{\mathrm{MMI,MMII}}$ between the diel call rate time series of two distinct MM species is calculated via,

$$r_{MM_I,MM_{II}}=\frac{\sum_{k=1}^{N}(c_{MM_I}(t_k)-\overline{c_{MM_I}})(c_{MM_{II}}(t_k)-\overline{c_{MM_{II}}})}{\sqrt{\sum_{k=1}^{N}(c_{MM_I}(t_k)-\overline{c_{MM_I}})^2}\sqrt{\sum_{k=1}^{N}(c_{MM_{II}}(t_k)-\overline{c_{MM_{II}}})^2}} \qquad (2)$$

where the samples at time t_j and t_k are independent for $j \neq k$, so that $t_{k+1}-t_k=1.25$ h.

The cumulative MM call rate distributions in Fig. 3b and Extended Data Fig. 8 are plotted as functions of decreasing distance from shoaling herring during night and day respectively, and so take the value 0 at long ranges from herring shoaling locations and monotonically increase to 1 at herring shoaling locations. The probability density function (PDF) for MM call rate density as a function of range from shoaling herring can be obtained as the absolute value of the range—derivative of the cumulative call rate distribution. The *e*-folding decay range of the cumulative call rate distribution for each MM species is the distance from herring shoals where the cumulative call rate distribution decays to $1/e=0.37$, so that 63% of vocalizations from that species are contained within the *e*-folding decay range.

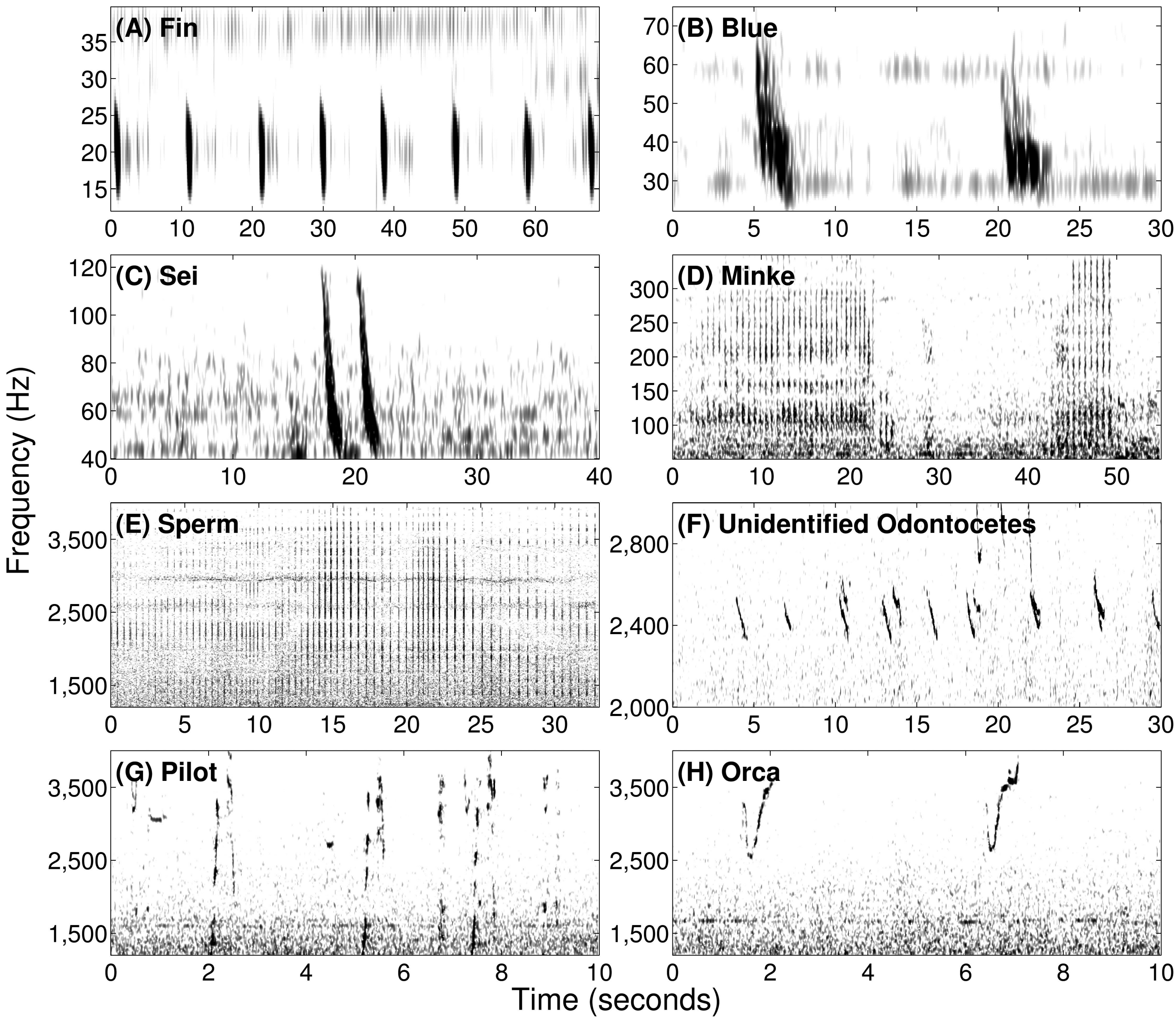
On the high temporal correlation between blue whale vocalization rate and herring shoaling density. Besides the fish-feeding hypothesis, another possible explanation is that both blue whales and herring are responding to a common environmental stimulus, such as changing light level or their common prey-zooplankton abundance. However, the annual biological sampling of the autumn season spawning herring over multiple years, including this 2006 experiment time period, indicates that the herring have largely empty stomachs and are generally not feeding while engaged in spawning activities[15].

Previous studies of diel dependence of blue vocalizations in their zooplankton-feeding areas found higher call rates in the night than day for longer duration (approximately 20 s) type B calls, associated with mating and social interaction, and proposed an inverse relationship between call rate and level of feeding activity[27]. It is also possible that the type D calls measured here are related to increased social interaction between blue whales at night.

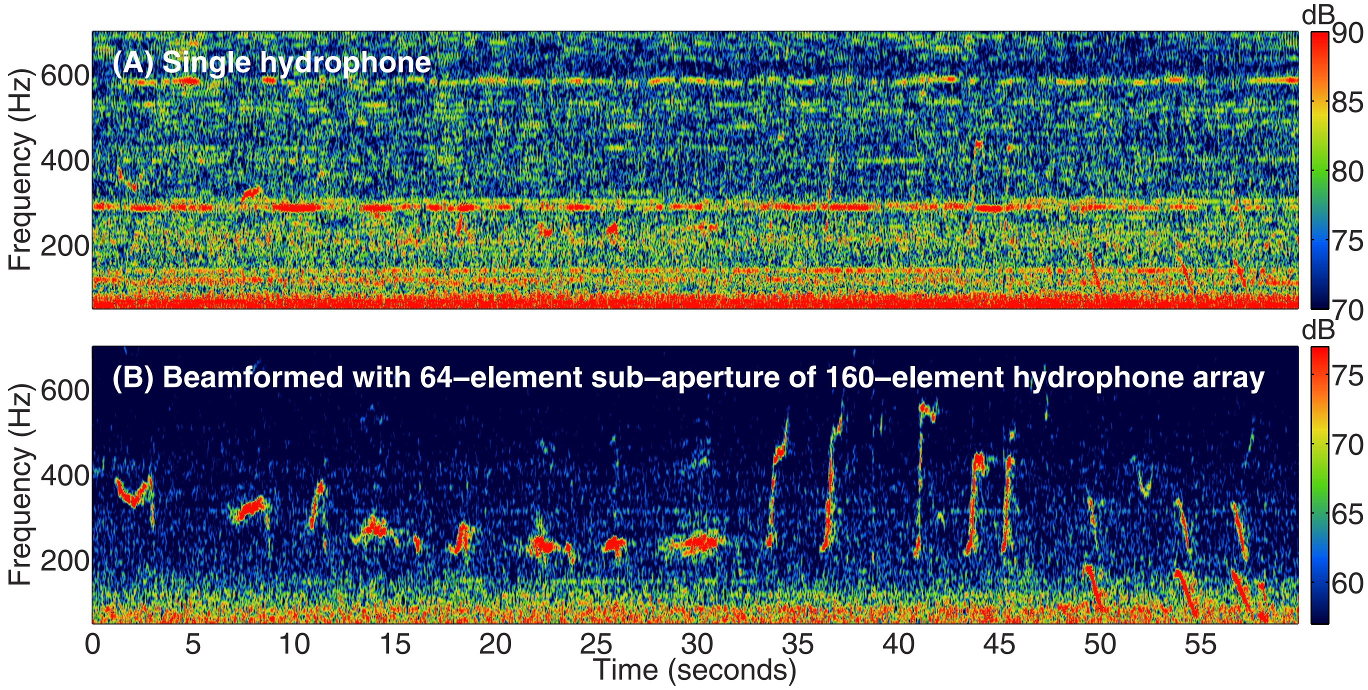
The Supplementary Information provides more details on marine mammal, fish and zooplankton distributions in GOM and comparison to other ocean environments.

31. Becker, K. & Preston, J. R. The ONR five octave research array (FORA) at Penn State. *Proc. OCEANS '03* **5**, 26072610 (2003).
32. Overholtz, W. J. *et al.* Stock assessment of the Gulf of Maine-Georges Bank Atlantic herring complex, 2003. *Northeast Fisheries Science Center References Document* **4,** 1–290 (2004).
33. Melvin, G. D. & Stephenson, R. L. The dynamics of a recovering fish stock: Georges Bank herring. *ICES J. Mar. Sci.* **64,** 69–82 (2007).
34. Read, A. J. & Brownstein, C. R. Considering other consumers: fisheries, predators, and Atlantic herring in the Gulf of Maine. *Conserv. Ecol.* **7,** 2 (2003).
35. Oppenheim, A. V., Willsky, A. S. & Nawab, S. H. *Signals and Systems* (PrenticeHall, 1997).
36. Johnson, D. H. & Dudgeon, D. E. *Array Signal Processing: Concepts and Techniques* (Simon & Schuster, 1992).
37. Baumgartner, M. F. *et al.* A generalized baleen whale call detection and classification system. *J. Acoust. Soc. Am.* **129,** 2889–2902 (2011).
38. Shapiro, A. D. & Wang, C. A versatile pitch tracking algorithm: From human speech to killer whale vocalizations. *J. Acoust. Soc. Am.* **126,** 451–459 (2009).
39. Wang, C. & Seneff, S. Robust pitch tracking for prosodic modeling in telephone speech. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Piscataway, NJ) 1143–1145 (2000).
40. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification* 2nd edn (John Wiley & Sons, 2001).

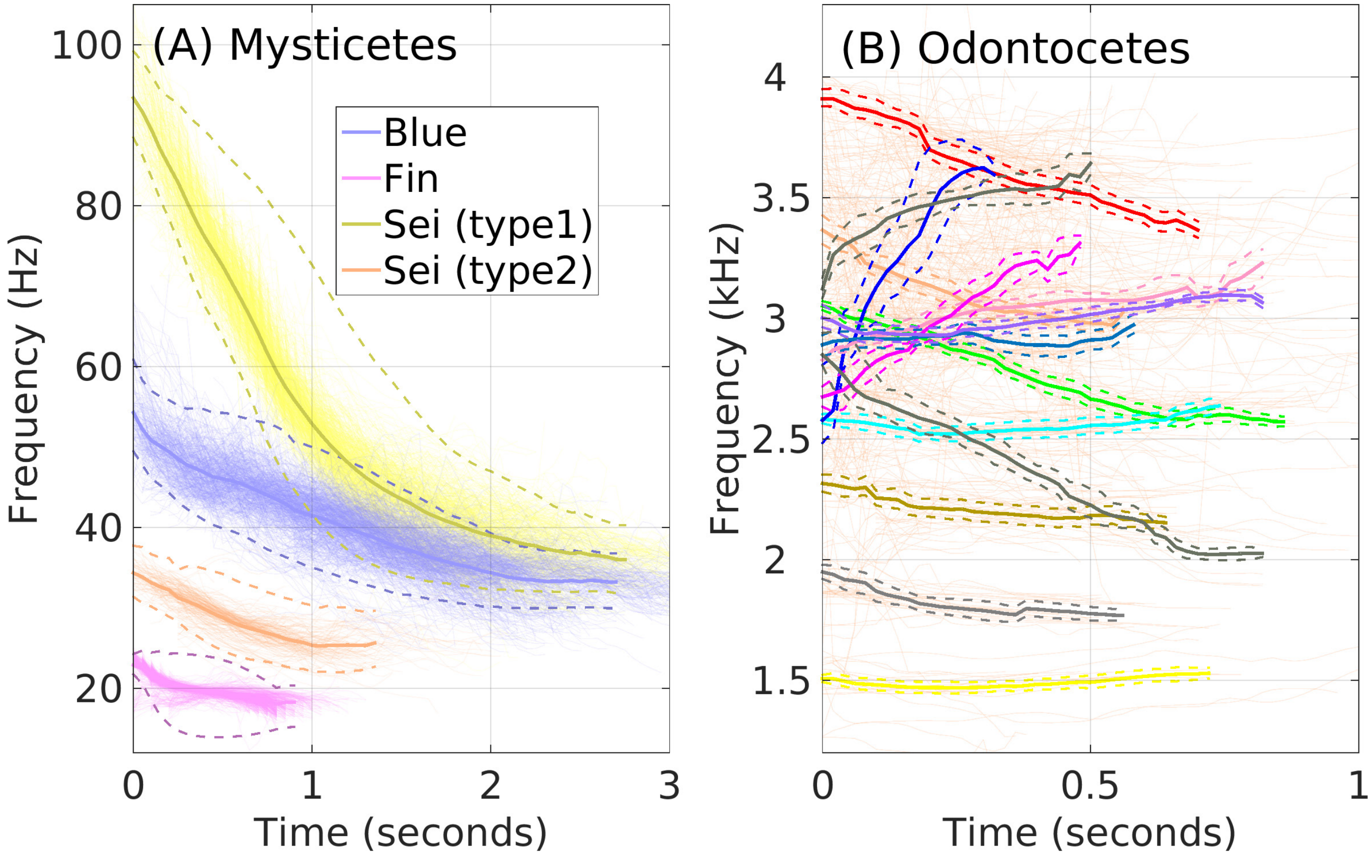
41. Watkins, W. A., Tyack, P., Moore, K. E. & Bird, J. E. The 20-Hz signals of finback whales (*Balaenoptera physalus*). *J. Acoust. Soc. Am.* **82,** 1901–1912 (1987).
42. Edds, P. L. Characteristics of finback *Balaenoptera physalus* vocalizations in the St. Lawrence Estuary. *Bioacoustics* **1,** 131–149 (1988).
43. Nieukirk, S. L. *et al.* Sounds from airguns and fin whales recorded in the mid-Atlantic Ocean, 1999–2009. *J. Acoust. Soc. Am.* **131,** 1102–1112 (2012).
44. Thompson, P. O., Findley, L. T. & Vidal, O. 20 Hz pulses and other vocalizations of fin whales, *Balaenoptera physalus*, in the Gulf of California, Mexico. *J. Acoust. Soc. Am.* **92,** 3051–3057 (1992).
45. McDonald, M. A., Hildebrand, J. A. & Webb, S. C. Blue and fin whales observed on a seafloor array in the Northeast Pacific. *J. Acoust. Soc. Am.* **98,** 712–721 (1995).
46. Croll, D. A. *et al.* Only male fin whales sing loud songs. *Nature* **417,** 809 (2002).
47. Oleson, E. M. Behavioral context of call production by eastern North Pacific blue whales. *Mar. Ecol. Prog. Ser.* **330,** 269–284 (2007).
48. McDonald, M. A., Calambokidis, J., Teranishi, A. M. & Hildebrand, J. A. The acoustic calls of blue whales off California with gender data. *J. Acoust. Soc. Am.* **109,** 1728–1735 (2001).
49. Thode, A. M., D'Spain, G. L. & Kuperman, W. A. Matched-field processing, geoacoustic inversion, and source signature recovery of blue whale vocalizations. *J. Acoust. Soc. Am.* **107,** 1286–1300 (2000).
50. Di Iorio, L. & Clark, C. W. Exposure to seismic survey alters blue whale acoustic communication. *Biol. Lett.* **6,** 51–54 (2010).
51. Rankin, S. & Barlow, J. Vocalizations of the sei whale *Balaenoptera borealis* off the Hawaiian Islands. *Bioacoustics* **16,** 137–145 (2007).
52. Baumgartner, M. F. *et al.* Low frequency vocalizations attributed to sei whales (*Balaenoptera borealis*). *J. Acoust. Soc. Am.* **124,** 1339–1349 (2008).
53. Edds-Walton, P. L. Vocalizations of minke whales *Balaenoptera acutorostrata* in the St. Lawrence Estuary. *Bioacoustics* **11,** 31–50 (2000).
54. Mellinger, D. K., Carson, C. D. & Clark, C. W. Characteristics of minke whale (*Balaenoptera acutorostrata*) pulse trains recorded near Puerto Rico. *Mar. Mamm. Sci.* **16,** 739–756 (2000).
55. Risch, D. *et al.* Minke whale acoustic behavior and multi-year seasonal and diel vocalization patterns in Massachusetts Bay, USA. *Mar. Ecol. Prog. Ser.* **489,** 279–295 (2013).
56. Payne, R. S. & McVay, S. Songs of humpback whales. *Science* **173,** 585–597 (1971).
57. Cerchio, S. & Dahlheim, M. Variation in feeding vocalizations of humpback whales (*Megaptera novaeangliae*) from southeast Alaska. *Bioacoustics* **11,** 277–295 (2001).
58. Cato, D., McCauley, R., Rogers, T. & Noad, M. Passive acoustics for monitoring marine animals-progress and challenges. *Proceedings of Acoustics* **2006,** 453–460 (2006).
59. Mattila, D. K., Guinee, L. N. & Mayo, C. A. Humpback whale songs on a North Atlantic feeding ground. *J. Mamm.* **68,** 880–883 (1987).
60. McSweeney, D. J., Chu, K. C., Dolphin, W. F. & Guinee, L. N. North Pacific humpback whale songs: a comparison of southeast Alaskan feeding ground songs with Hawaiian wintering ground songs. *Mar. Mamm. Sci.* **5,** 139–148 (1989).
61. Clark, C. W. & Clapham, P. J. Acoustic monitoring on a humpback whale (*Megaptera novaeangliae*) feeding ground shows continual singing into late spring. *Proc. Royal Soc. B* **271,** 1051–1057 (2004).
62. Watkins, W. A. & Schevill, W. E. Sperm whale codas. *J. Acoust. Soc. Am.* **62,** 1485–1490 (1977).
63. Tiemann, C. O., Thode, A. M., Straley, J., O'Connell, V. & Folkert, K. Three-dimensional localization of sperm whales using a single hydrophonea. *J. Acoust. Soc. Am.* **120,** 2355–2365 (2006).
64. Nemiroff, L. & Whitehead, H. Structural characteristics of pulsed calls of long-finned pilot whales *Globicephala melas*. *Bioacoustics* **19,** 67–92 (2009).
65. Weilgart, L. S. & Whitehead, H. Vocalizations of the North Atlantic pilot whale (*Globicephala melas*) as related to behavioral contexts. *Behav. Ecol. Sociobiol.* **26,** 399–402 (1990).
66. Tyson, R. B., Nowacek, D. P. & Miller, P. J. Nonlinear phenomena in the vocalizations of North Atlantic right whales (*Eubalaena glacialis*) and killer whales (*Orcinus orca*). *J. Acoust. Soc. Am.* **122,** 1365–1373 (2007).
67. Simon, M., Ugarte, F., Wahlberg, M. & Miller, L. A. Icelandic killer whales *Orcinus orca* use a pulsed call suitable for manipulating the schooling behaviour of herring *Clupea harengus*. *Bioacoustics* **16,** 57–74 (2006).
68. Filatova, O. A. *et al.* Call diversity in the North Pacific killer whale populations: implications for dialect evolution and population history. *Anim. Behav.* **83,** 595–603 (2012).
69. Oliveira, C., Wahlberg, M., Johnson, M., Miller, P. J. & Madsen, P. T. The function of male sperm whale slow clicks in a high latitude habitat: communication, echolocation, or prey debilitation? *J. Acoust. Soc. Am.* **133,** 3135–3144 (2013).
70. Ford, J. K. Acoustic behavior of resident killer whales (*Orcinus orca*) off Vancouver Island, British Columbia. *Can. J. Zool.* **67,** 727–745 (1989).
71. Gong, Z., Tran, D. & Ratilal, P. Comparing passive source localization and tracking approaches with a towed horizontal receiver array in an ocean waveguide. *J. Acoust. Soc. Am.* **134,** 3705–3720 (2013).
72. Lee, S. & Makris, N. C. The array invariant. *J. Acoust. Soc. Am.* **119,** 336–351 (2006).
73. Gong, Z., Ratilal, P. & Makris, N. C. Simultaneous localization of multiple broadband non-impulsive acoustic sources in an ocean waveguide using the array invariant. *J. Acoust. Soc. Am.* **138,** 2649–2667 (2015).
74. Tran, D., Andrews, M. & Ratilal, P. Probability distribution for energy of saturated broadband ocean acoustic transmission: results from Gulf of Maine 2006 experiment. *J. Acoust. Soc. Am.* **132,** 3659–3672 (2012).
75. Jain, A., Ignisca, A., Yi, D., Ratilal, P. & Makris, N. Feasibility of ocean acoustic waveguide remote sensing (OAWRS) of Atlantic cod with seafloor scattering limitations. Remote Sens. **6,** 180–208 (2014).
76. Jagannathan, S., Kusel, E., Ratilal, P. & Makris, N. Scattering from extended targets in range-dependent fluctuating ocean-waveguides with clutter from theory and experiments. *J. Acoust. Soc. Am.* **132,** 680–693 (2012).
77. Andrews, M., Chen, T. & Ratilal, P. Empirical dependence of acoustic transmission scintillation statistics on bandwidth, frequency and range in New Jersey continental shelf. *J. Acoust. Soc. Am.* **125,** 111–124 (2009).
78. Collins, M. D. Generalization of the split-step Padé solution. *J. Acoust. Soc. Am.* **93,** 1736–1742 (1983).
79. Ainslie, M. A. Neglect of bandwidth of Odontocetes echo location clicks biases propagation loss and single hydrophone population estimates. *J. Acoust. Soc. Am.* **134,** 3506–3512 (2013).
80. Makris, N. C. The effect of saturated transmission scintillation on ocean-acoustic intensity measurements. *J. Acoust. Soc. Am.* **100,** 769–783 (1996).
81. Bertsatos, I., Zanolin, M., Chen, T. R., Ratilal, P. & Makris, N. C. General second order covariance of Gaussian maximum likelihood estimate applied to passive source localization in a fluctuating ocean waveguide. *J. Acoust. Soc. Am.* **128,** 2635–2651 (2010).
82. Thode, A. *et al.* Necessary conditions for a maximum likelihood estimate to become asymptotically unbiased and attain the Cramer-Rao lower bound Part II: range and depth localization of a sound source in an ocean waveguide. *J. Acoust. Soc. Am.* **112,** 1890–1910 (2002).
83. Andrews, M., Gong, Z. & Ratilal, P. Effects of multiple scattering, attenuation and dispersion in waveguide sensing of fish. *J. Acoust. Soc. Am.* **130,** 1253–1271 (2011).
84. DiFranco, J. V. & Rubin, W. L. *Radar Detection (Artech House, 1980).*
85. Bergmann, P. G., Yaspan, A., Gerjuoy, E., Major, J. K. & Wildt, R. *Physics of Sound in the Sea* (Gordon and Breach, 1968).
86. Urick, R. J. *Principles of Underwater Sound* 29–65 and 343–366 (McGraw Hill, 1983).
87. Širović, A., Hildebrand, J. A. & Wiggins, S. M. Blue and fin whale call source levels and propagation range in Southern Ocean. *J. Acoust. Soc. Am.* **122,** 1208–1215 (2007).
88. Weirathmueller, M. J., Wilcock, W. S. & Soule, D. C. Source levels of fin whale 20 Hz pulses measured in the Northeast Pacific Ocean. *J. Acoust. Soc. Am.* **133,** 741–749 (2013).
89. Newhall, A. E., Lin, Y. T., Lynch, J. F., Baumgartner, M. F. & Gawarkiewicz, G. G. Long distance passive localization of vocalizing sei whales using an acoustic normal mode approach. *J. Acoust. Soc. Am.* **131,** 1814–1825 (2012).
90. Jensen, F. B., Kuperman, W. A., Porter, M. B. & Schmidt, H. *Computational Ocean Acoustics* 708–713 (Springer-Verlag, 2011).
91. Clay, C. S. & Medwin, H. *Acoustical Oceanography* 494–501 (John Wiley, 1977).
92. Burdic, W. S. *Underwater Acoustic System Analysis* 322–360 (Prentice-Hall, 1984).
93. Küsel E. T. *et al.* Cetacean population density estimation from single fixed sensors using passive acoustics. *J. Acoust. Soc. Am.* **129,** 3610–3622 (1983).
94. Marques, T. A., Thomas, L., Ward, J., DiMarzio, N. & Tyack, P. L. Estimating cetacean population density using fixed passive acoustic sensors: an example with Blainvilles beaked whales. *J. Acoust. Soc. Am.* **125,** 1982–1994 (2009).
95. Barlow, J. & Taylor, B. L. Estimates of sperm whale abundance in the northeastern temperate Pacific from a combined acoustic and visual survey. *Mar. Mamm. Sci*. **21,** 429–445 (2005).
96. Martin, S. W. *et al.* Estimating minke whale (*Balaenoptera acutorostrata*) boing sound density using passive acoustic sensors. *Mar. Mamm. Sci*. **29,** 142–158 (2013).
97. Marques, T. A., Munger, L., Thomas, L., Wiggins, S. & Hildebrand, J. A. Estimating North Pacific right whale *Eubalaena japonica* density using passive acoustic cue counting. Endanger. *Species Res.* **13,** 163–172 (2011).
98. Marques, T. A. *et al.* Estimating animal population density using passive acoustics. *Biol. Rev. Camb. Philos. Soc.* **88,** 287–309 (2013).
99. Battista, T. A., Clark, R. D. & Pittman, S. An ecological characterization of the Stellwagen Bank national marine sanctuary region: oceanographic, biogeographic, and contaminants assessment. US Department of Commerce, National Oceanic and Atmospheric Administration, National Ocean Service, Center for Coastal Monitoring and Assessment, 265282 (2006).
100. Stark, H. & Woods, J. *Probability, Statistics, and Random Processes for Engineers* (Prentice Hall, 2011).



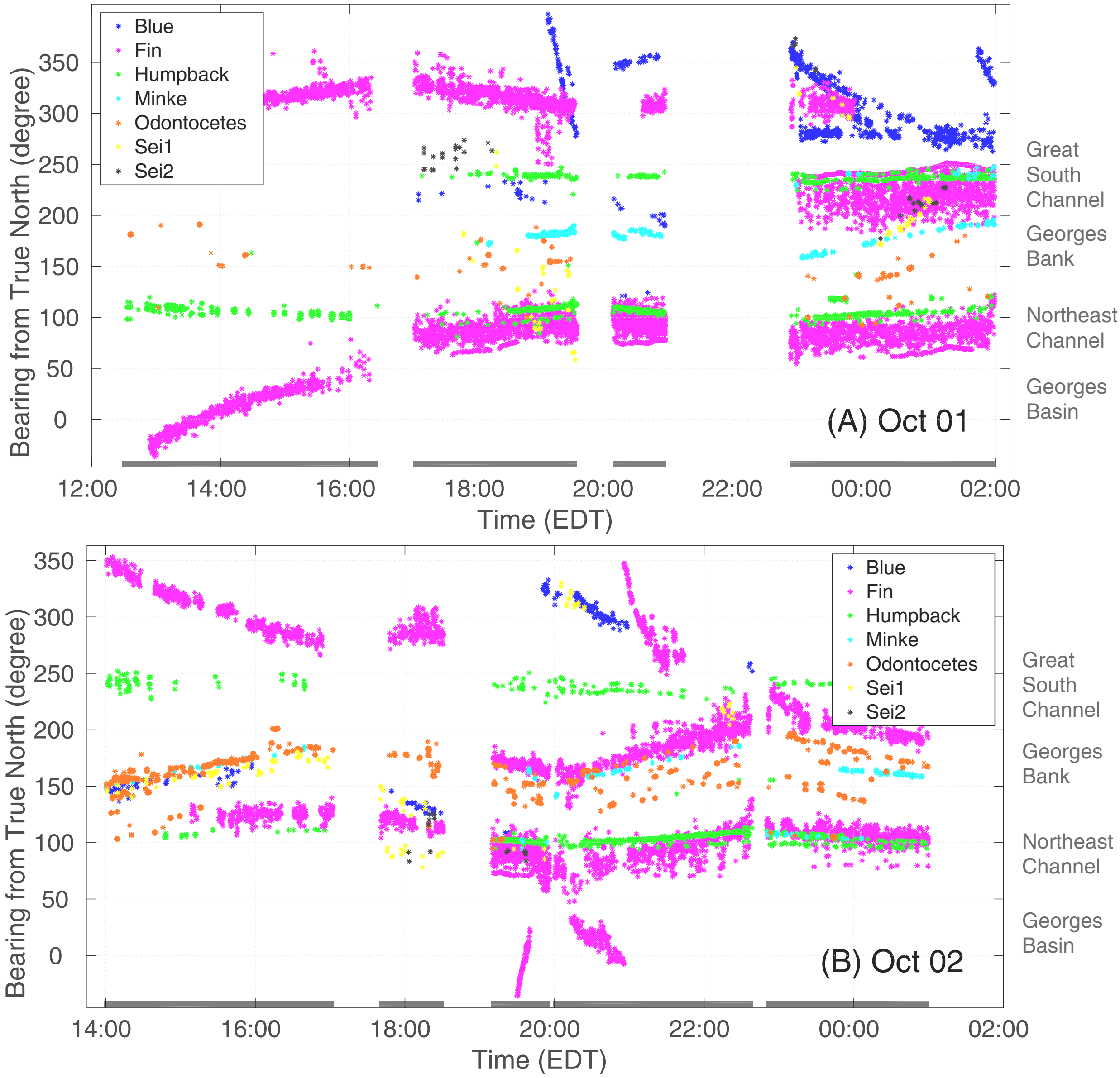
Extended Data Figure 1 | Spectrograms of MM vocalizations. a–h, Beamformed spectrograms of typical repetitive vocalizations from diverse MM species observed using the POAWRS receiver array in the Gulf of Maine from 19 September to 6 October 2006.



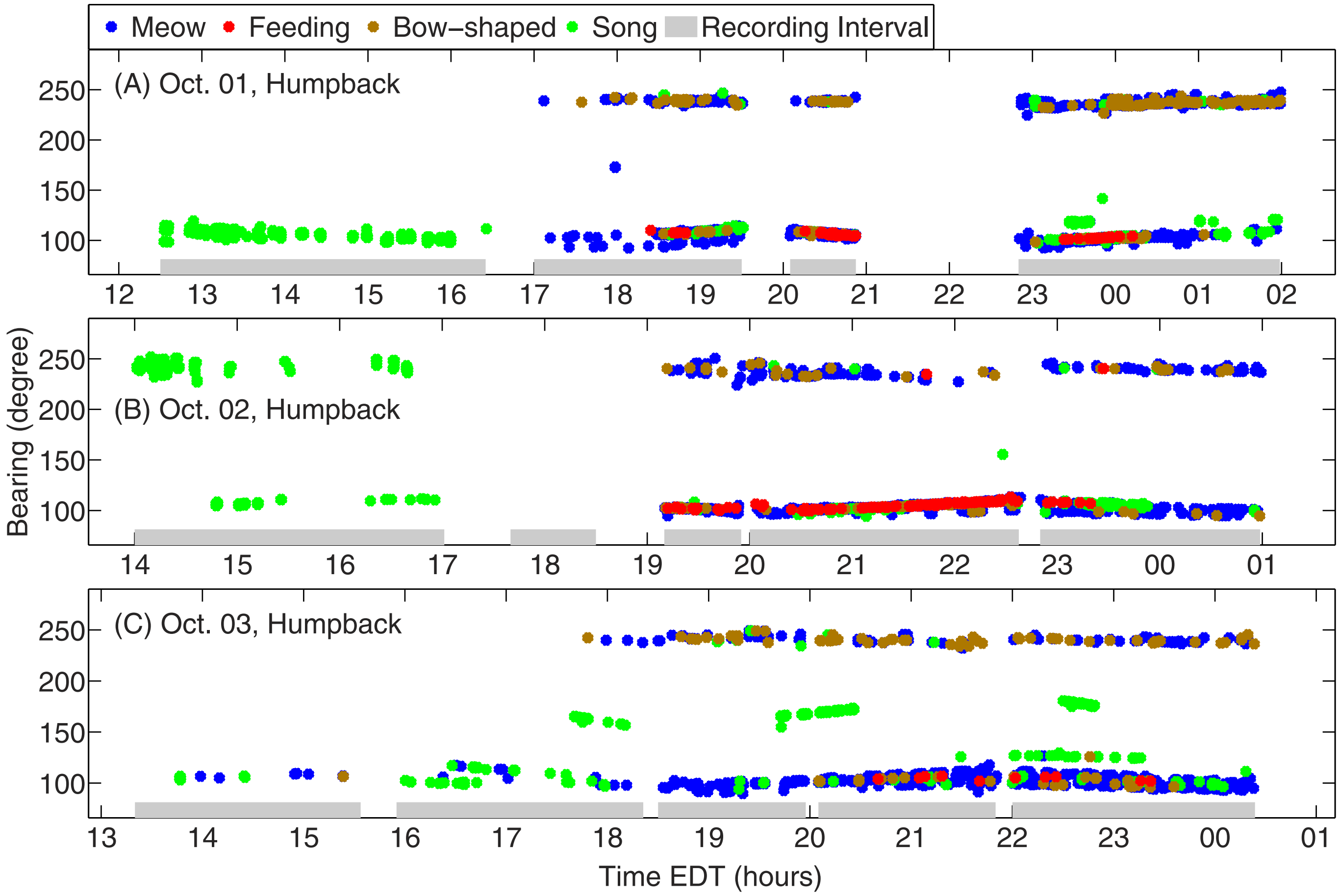
Extended Data Figure 2 | Coherent array processing enhances SNR. **a**, **b**, Compare single hydrophone measured spectrogram (**a**) with spectrogram after coherent beamforming (**b**) with 64-element sub-aperture of POAWRS 160-element hydrophone array. The song vocalization from a humpback individual roughly 35 km away from the POAWRS receiver array recorded on 2 October 2006 at 23:48:45 EDT is enhanced by 18 dB above the background noise after beamforming in **b** where whale bearing is −64.16° from array broadside.



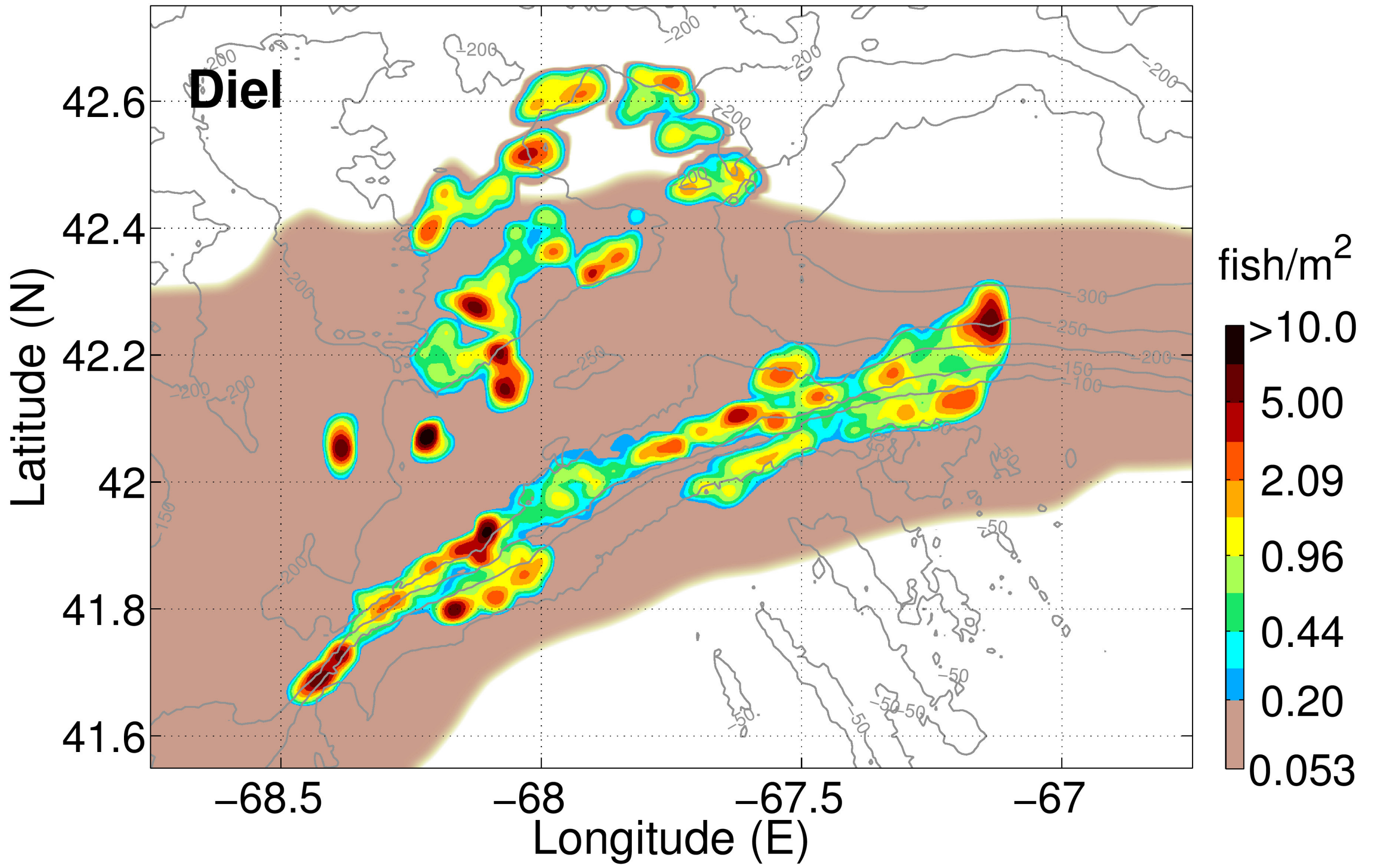
Extended Data Figure 3 | Pitch-tracks of baleen and toothed whales. **a**, Pitch-tracks of repetitive mysticete vocalizations in the 10 to 100 Hz range. Thick solid curves are the means of roughly 500 to 1,000 vocalizations of each type. Mean instantaneous bandwidth of the pitch-tracks are indicated by the dashed curve. Even though blue and sei type I calls have some overlapping bandwidth, they can be well separated using the upper frequency f_U and slope $df/d\tau$ features (Extended Data Table 2). **b**, Mean pitch-track and instantaneous bandwidth of repetitive odontocete downsweep vocalizations in the 1 to 4 kHz range.



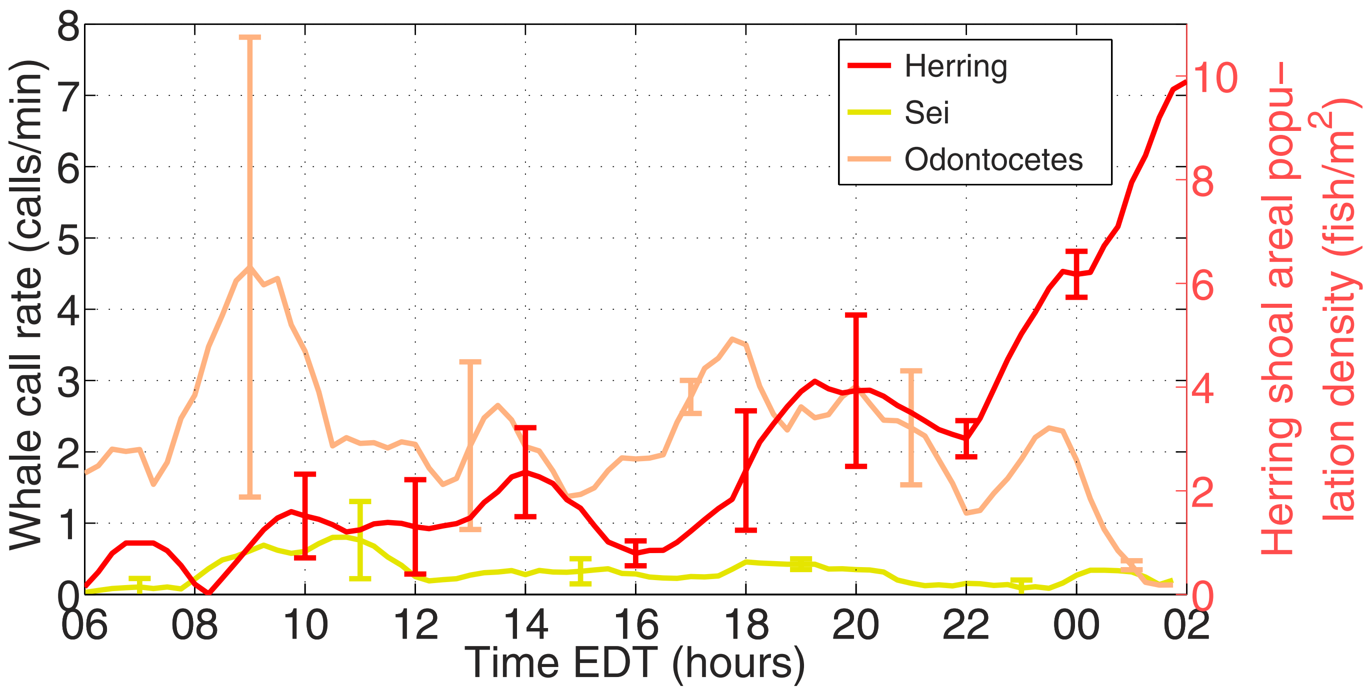
Extended Data Figure 4 | Daily POAWRS measured MM vocalization bearings. a, **b**, MM vocalization bearings from diverse species measured by POAWRS receiver array on 1 October 2006 (**a**) and 2 October 2006 (**b**). The bearings are measured from true North in clockwise direction with respect to the instantaneous spatial locations of the receiver array centre. The techniques used here for resolving source bearing ambiguity about the horizontal line-array axis are provided in Methods section 3. The shaded bars on the *x* axis indicate the operation time periods of the receiver array.



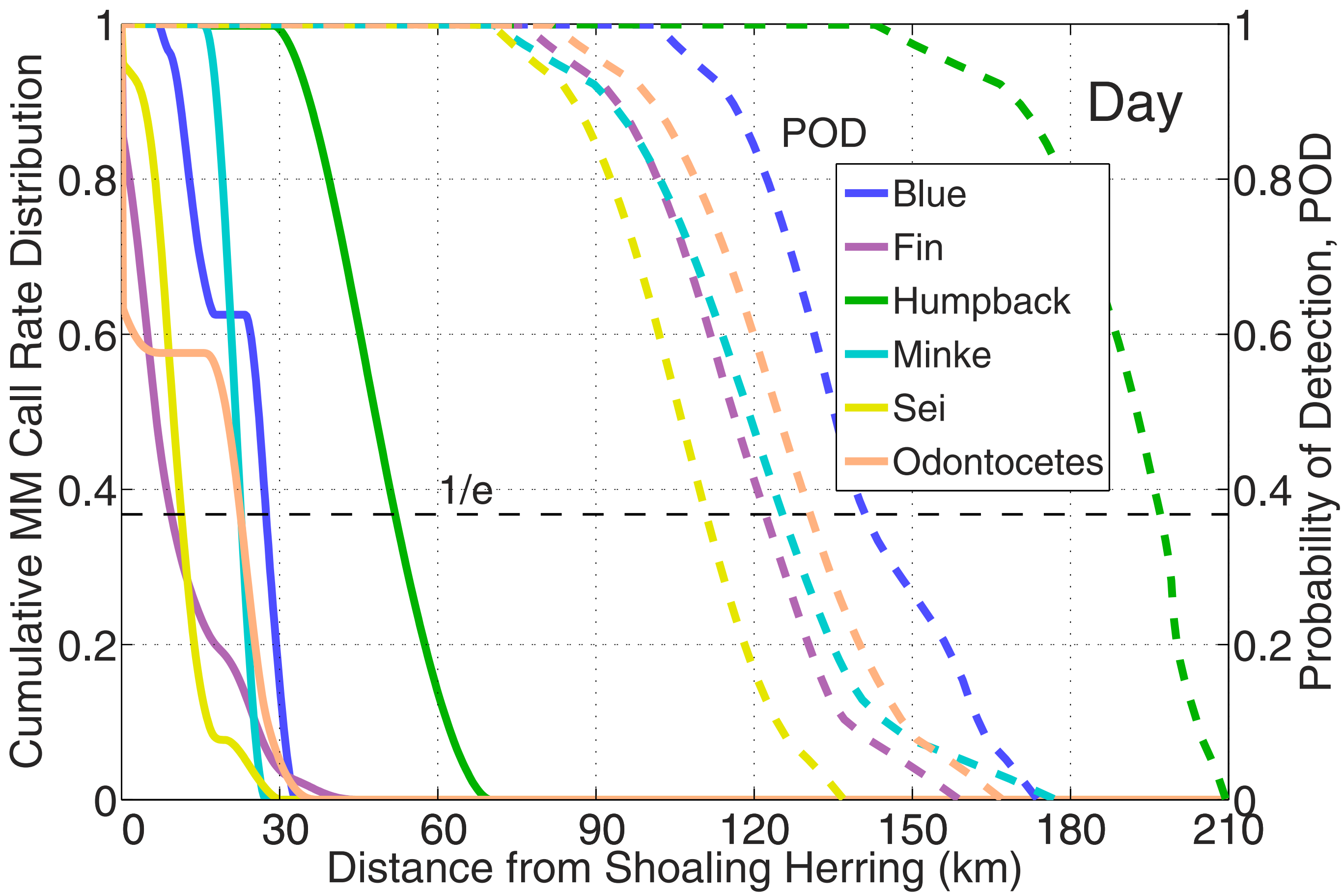
Extended Data Figure 5 | Daily humpback vocalization repertoire. a–c, Bearings and repertoire of humpback vocalizations measured by POAWRS receiver array on 1 October 2006 (**a**), 2 October 2006 (**b**), and 3 October 2006 (**c**). The 'meow', 'bow', and 'feeding' call characteristics are provided in ref. 7.



Extended Data Figure 6 | Diel Atlantic herring shoaling areal population densities. Measured herring shoaling areal population densities (ranging from 0.2 fish per m^2 to over 10 fish per m^2) determined from OAWRS[12,13] survey in the Gulf of Maine from 26 September 2006 to 6 October 2006, upon calibration with tens of thousands of coincident and co-located conventional ultrasonic fisheries echosounding measurements, combined with trawl sampling for identity and biological–physiological characterization of fish populations[15,16]. The mean diffuse herring density of 0.053 fish per m^2 is determined from conventional ultrasonic fisheries echosounding. The bathymetric data (contours shown in grey) were obtained from the US National Centers for Environmental Information.



Extended Data Figure 7 | Diel MM call rate and herring shoal areal population density time series. Mean diel call rates for sei whales and odontocetes in general are not correlated to the diel Atlantic herring shoal mean areal population density. The error bars indicate standard deviations obtained from averaging the time series over multiple diel cycles from 26 September 2006 to 6 October 2006. The period from roughly 2–6 EDT contains a data gap.



Extended Data Figure 8 | Cumulative diurnal MM call rate distribution. Cumulative diurnal MM vocalization rate distribution and azimuthally-averaged POAWRS MM POD as a function of minimum distance from diurnal herring shoaling areas. The *e*-folding distances of the cumulative MM vocalization rate distributions decrease from day (shown here) to night (in Fig. 3b) by 27.3 to 7 km (blue), 9.3 to 3.9 km (fin), 51.7 to 3.5 km (humpback), 22.5 to 0 km (minke), 11.2 to 8.1 km (sei), and 22.4 to 5.5 km (odontocetes). The percentage of vocalizations that fully overlap with herring shoaling areas increase from day to night by 0% to 18% (blue), 14% to 40% (fin), 6% to 44% (humpback), 0% to 71% (minke), and 5% to 24% (sei), but decrease by 36% to 29% (odontocetes).

Extended Data Table 1 | MM species daily call rate and temporal correlations

MM species	C_{MM} calls/day	$\frac{C_{MM,\text{night}}}{C_{MM,\text{day}}}$	$r_{MM,fish}$	$r_{MM_I,MM_{II}}$					
				blue	fin	humpback	minke	sei	odontocetes
blue	470	10	0.91	1	0.67	0.78	0.48	-0.17	-0.46
fin	14,000	2.5	0.82	0.67	1	0.85	0.79	0.02	-0.27
humpback	2,000	10	0.87	0.78	0.85	1	0.84	-0.29	-0.30
minke	690	5	0.64	0.48	0.79	0.84	1	-0.18	-0.20
sei	440	0.77	-0.11	-0.17	0.02	-0.29	-0.18	1	0.44
odontocetes	3200	0.83	-0.42	-0.46	-0.27	-0.30	-0.20	0.44	1

The temporal correlation $r_{\text{MM,fish}}$ of MM vocalization rate time series to fish shoaling areal population density time series, as well as the temporal correlation $r_{\text{MMI, MMII}}$ between distinct MM species vocalization rate time series over the diel cycle are calculated using Methods equations (1) and (2) respectively.

Extended Data Table 2 | Large baleen whale repetitive vocalization pitch-track features

Characteristics	fin	blue	sei (type 1)	sei (type 2)
f_L (Hz)	13.7 ± 0.6	28 ± 3	32 ± 6	22 ± 3
f_U (Hz)	24.9 ± 1	56 ± 8	88 ± 12	36 ± 4
$\overline{f}$ (Hz)	19.8 ± 0.4	40 ± 4	50 ± 7	28 ± 3
$\overline{B}$ (Hz)	8 ± 1	12.5 ± 4	21 ± 5	8.1 ± 1.5
$\overline{B}/\overline{f}$	0.40 ± 0.05	0.31 ± 0.08	0.4 ± 0.1	0.29 ± 0.06
τ (s)	0.8 ± 0.2	2 ± 0.5	2 ± 0.5	1 ± 0.3
$\frac{df}{d\tau}$ (Hz/s)	-4.6 ± 2	-9 ± 5	-24 $\pm$ 10	-6.6 ± 3
$\frac{d^2 f}{d\tau^2}$ (Hz/s^2)	2.5 ± 9	0.5 ± 6	13 ± 8	5.6 ± 8

Vocalization characteristics in the 10 to 100 Hz range from large baleen whales detected with POAWRS receiver array in the Gulf of Maine in autumn 2006. The characteristics are the lower f_L, upper f_U, and mean $\bar{f}$ frequencies, mean instantaneous bandwidth $\bar{B}$, relative instantaneous bandwidth $\bar{B}/\bar{f}$ duration τ, slope $df/d\tau$, and curvature $d^2f/d\tau^2$. The slope and curvature are obtained from second order nonlinear curve-fit to the vocalization traces obtained via pitch-tracking.

doi:10.1038/nature17187

Sensory experience regulates cortical inhibition by inducing IGF1 in VIP neurons

A. R. Mardinly[1]*, I. Spiegel[2]*, A. Patrizi[3], E. Centofante[3], J. E. Bazinet[2], C. P. Tzeng[2], C. Mandel-Brehm[2], D. A. Harmin[2], H. Adesnik[1], M. Fagiolini[3] & M. E. Greenberg[2]

Inhibitory neurons regulate the adaptation of neural circuits to sensory experience[1], but the molecular mechanisms by which experience controls the connectivity between different types of inhibitory neuron[2,3] to regulate cortical plasticity are largely unknown. Here we show that exposure of dark-housed mice to light induces a gene program in cortical vasoactive intestinal peptide (VIP)-expressing neurons that is markedly distinct from that induced in excitatory neurons and other subtypes of inhibitory neuron. We identify *Igf1* as one of several activity-regulated genes that are specific to VIP neurons, and demonstrate that IGF1 functions cell-autonomously in VIP neurons to increase inhibitory synaptic input onto these neurons. Our findings further suggest that in cortical VIP neurons, experience-dependent gene transcription regulates visual acuity by activating the expression of IGF1, thus promoting the inhibition of disinhibitory neurons[3–5] and affecting inhibition onto cortical pyramidal neurons.

To explore how sensory experience affects gene expression in VIP neurons, we examined this process in the visual cortex of adult mice that were housed in standard conditions, in complete darkness (that is, dark-housed), or dark-housed and then exposed to light for increasing amounts of time[6,7] (Fig. 1a). Light deprivation for as little as 12 h drives robust gene expression after light exposure, and increasing durations of dark-housing accentuate the gene induction response (Extended Data Fig. 1a) irrespective of the phase of the circadian rhythm (Extended Data Fig. 1b). To purify RNA selectively from VIP-expressing and other inhibitory neuron subtypes, we generated mice that were heterozygous for alleles of either *Vip-cre*, *Sst-cre* or *Pv-cre*, and were also heterozygous for the *Rpl22-HA* (RiboTag) allele[8], which expresses a haemagglutinin (HA)-tagged ribosomal protein specifically in Cre-expressing neurons (Fig. 1a). For purposes of comparison, we also purified ribosome-bound RNA from excitatory and inhibitory neurons, labelled by *Emx1-cre* or *Gad2-cre*.

By quantitative real-time PCR (qPCR), we find that messenger RNAs for cell-type-specific marker genes are highly enriched in the appropriate samples (Extended Data Fig. 1c) and that light exposure induces the expression of early-response genes in each Cre line (Extended Data Fig. 1d). To quantify experience-induced gene expression at a genome-wide level, we performed RNA-seq on RNA isolated from the dark-housed/light-exposed RiboTag-mice (Supplementary Table 1) (Fig. 1b, c and Extended Data Fig. 2a, b). This analysis identified genes which exhibited reproducible changes in expression levels in response to visual stimulation in at least one Cre line ($n = 602$; see Supplementary Table 2 and Methods) and thus allowed us to ask how levels of these experience-regulated genes are correlated across the different neuronal subtypes compared to non-regulated genes ($n = 13{,}678$) (Fig. 1d–i). We found that the expression of experience-regulated genes is remarkably dissimilar across different neuronal subtypes when compared to genes that are not regulated by sensory experience (irrespective of differences in the number or expression levels of experience-regulated genes; Fig. 1f–i and Extended Data Fig. 3a–d). While unique subsets of experience-responsive genes were identified in each neuronal subtype (Fig. 2a, b and Extended Data Fig. 3e, f), VIP neurons are the most responsive to sensory stimulation (Fig. 1d, e) and possess an experience-induced gene expression program that is markedly distinct from the other neuronal subtypes analysed (Fig. 1h, i). This suggests that in VIP neurons the experience-dependent gene program may have a unique function in adapting the cortex's neural circuits to sensory experience.

We hypothesized that experience-regulated genes that are specifically expressed and selectively regulated in VIP neurons are likely to have important functions in regulating the synaptic connectivity onto VIP neurons. Thus, we first identified the mRNAs that are specifically enriched in each subtype (Extended Data Fig. 4a; Methods) and cross-referenced these genes with the list of experience-regulated genes (Extended Data Fig. 4b). This analysis identified 31 genes that are both cell-type-specific and experience-regulated, 11 of which are specific to VIP neurons (Supplementary Table 4). Notably, secreted molecules are significantly over-represented in this gene set (GO-term 'Secreted' $P = 0.002$) and each type of neuron has its own set of cell-type-specific experience-regulated secreted factors, including four experience-induced secreted molecules that are specific to VIP neurons (*Igf1*, *Crh*, *Prok2*, *Fbln2*; Fig. 2b, Supplementary Table 4).

We next performed fluorescent *in situ* hybridization (FISH) on sections of visual cortices of dark-housed/light-exposed mice to quantify the percentage of cells that co-express an inhibitory subtype marker and the respective secreted factor (Fig. 2c–f). Of the four secreted factors, *Igf1* is the one factor that is expressed in the vast majority of VIP neurons, and whose expression is highly enriched in these neurons (Fig. 2d). We were unable to reliably identify *Fbln2*-expressing cells, and *Prok2* was expressed nearly exclusively in a sparse subpopulation of VIP neurons (Fig. 2f), consistent with the low expression level of these genes in the RiboTag-seq experiments (Extended Data Fig. 4c). While the FISH analysis revealed that in the cortex *Crh* is highly enriched in VIP neurons compared to PV and SST neurons (Fig. 2e), this gene is also expressed in *Pv-*/*Sst-*/*Vip*-negative cortical interneurons[9]. Since IGF1 is the sole experience-induced secreted factor that is selectively expressed in most VIP neurons, we focused our subsequent analysis on IGF1.

Previous reports have suggested that IGF1 is synthesized in the cortex[10], but the function of cortical-derived IGF1 was unknown. Global disruption of the *Igf1* gene results in abnormally small animals with smaller brains that contain smaller neurons with dendrites that are less branched and contain fewer synapses[11–13], and the effects of IGF1 on brain development and function are due at least in part to IGF1 that is produced by non-neural tissues and then enters the brain[14]. To investigate specifically VIP neuron-derived IGF1, we crossed *Vip-cre* mice to both IGF1 conditional-knockout mice[15] and Cre reporter mice. Disruption of *Igf1* specifically in VIP neurons had no effect on the

[1]Department of Molecular and Cellular Biology, University of California Berkeley, 205 Life Sciences Addition, Berkeley, California 94720, USA. [2]Department of Neurobiology, Harvard Medical School, 220 Longwood Ave, Boston, Massachusetts 02115, USA. [3]FM Kirby Neurobiology Center, Boston Children's Hospital, 3 Blackfan Circle, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

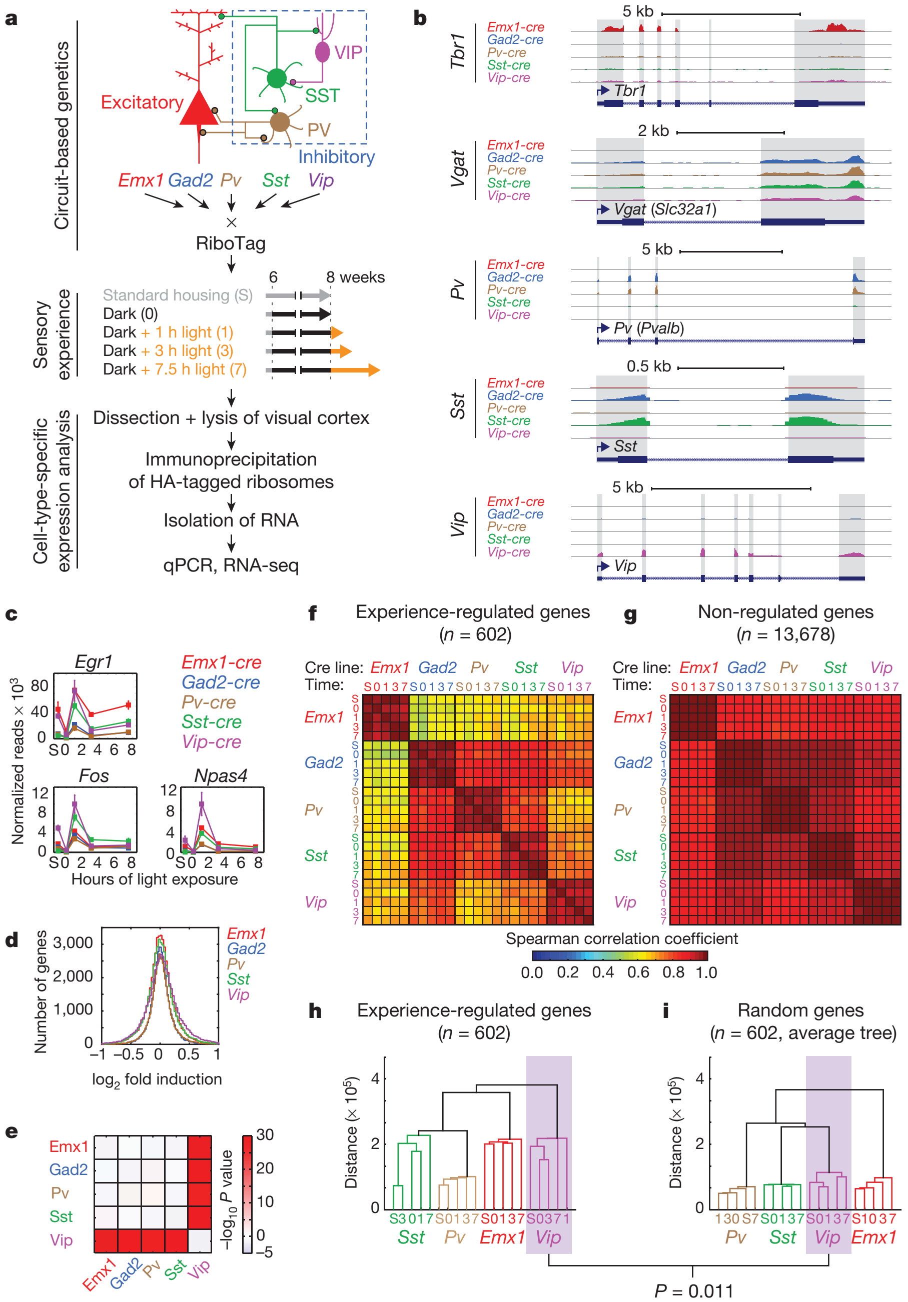


Figure 1 | VIP neurons mount a unique transcriptional response to sensory experience. a, Approach for purifying ribosome-bound RNA from Cre-expressing neurons in visual cortex following manipulation of visual experience. **b,** Representative RNA-seq tracks of cell-type-specific marker genes (exons shaded). **c,** Line plots showing the expression levels of selected early induced genes in each Cre line at each time point ($n = 3$, error bars, s.e.m). **d,** Histograms showing the distribution of all observed fold-changes for each Cre line (mean of three replicates). **e,** Colour-coded matrix showing the $-\log_{10}$ P value of pairwise corrected two-tailed t-tests of the fold-change distributions between each Cre line. **f, g,** Matrices of Spearman correlation coefficients computed from the expression levels of experience-regulated genes (**f**) or genes not regulated by experience (**g**) (mean of three replicates). **h, i,** Cladograms of all experience-regulated genes created by using the mean values of each gene in each sample (**h**) or the average cladogram created from 1,000 random sets of 602 random genes using the mean expression values of each gene in each sample (**i**) ($P = 0.011$, Monte Carlo test).

thickness of the cortical layers, on the number and layer distribution of VIP neurons, or on the size of VIP neuron cell bodies at postnatal day 21 (that is, P21) (Extended Data Fig. 5a–d). To test whether VIP neuron-derived IGF1 affects excitatory and/or inhibitory inputs to VIP neurons, we recorded miniature inhibitory or excitatory postsynaptic currents (mIPSCs or mEPSCs) in VIP neurons in acute visual cortex slices; we found that conditional deletion of *Igf1* in VIP neurons leads to a significant reduction in mIPSC frequency (Fig. 3a) but not amplitude (Fig. 3b). Since conditional deletion of *Igf1* had no effect on the frequency or amplitude of excitatory mEPSCs on VIP neurons (Fig. 3c, d), these findings suggest that VIP-neuron-derived IGF1 specifically enhances inhibitory synaptic input onto VIP neurons.

To test whether IGF1 functions cell-autonomously to regulate inhibitory input onto the cell from which it is expressed, we used a virus-based approach to acutely knockdown *Igf1* expression in only a few VIP neurons. We generated short hairpin RNA (shRNA) constructs against *Igf1* (Extended Data Fig. 6a, b), injected low titre AAVs expressing the shRNA and Cre-dependent enhanced green fluorescent protein (eGFP) into the visual cortex of P14–15 *Vip-cre* mice, and recorded mIPSCs and mEPSCs one week later (P20–P22) in eGFP-positive VIP neurons that are surrounded by non-infected VIP neurons (Fig. 3e). This sparse and acute knockdown of *Igf1* in VIP neurons using either of two distinct shRNAs against *Igf1* resulted in a marked reduction in mIPSC frequency and amplitude as compared to VIP neurons infected with a control shRNA (Fig. 3f, g), but had no significant effect on mEPSCs (Fig. 3h, i). These effects are not due to altered VIP neuron morphology (Fig. 3j and Extended Data Fig. 6c), indicating that VIP-neuron-derived IGF1 acutely promotes inhibition onto VIP neurons in a cell-autonomous manner.

To determine if VIP-neuron-derived IGF1 regulates inhibitory inputs onto other types of cortical neurons, we adopted a protocol that leads to widespread infection of neurons in the cortex (see Methods). Injecting AAVs into *Vip-*, *Pv-* or *Sst-cre* mice to label these cells with eGFP while knocking down *Igf1* in VIP neurons, we recorded mIPSCs from each cell type and found that early knockdown of *Igf1* in the cortex decreases mIPSCs frequency in VIP neurons, but does not affect mIPSCs onto

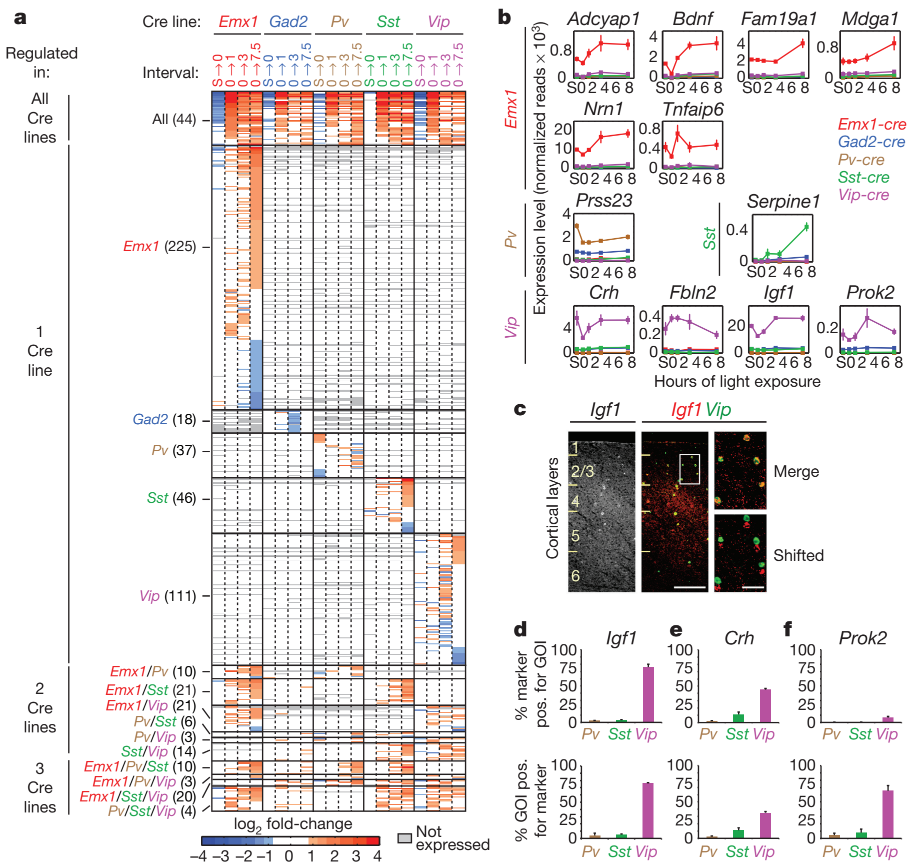


Figure 2 | IGF1 is an experience-induced cell-type-specific secreted factor in VIP neurons. a, Heat map showing the fold-change in expression of all experience-regulated genes across all stimulus intervals. **b,** Line plots of secreted factors that are experience-regulated and expressed in a cell-type-specific manner ($n = 3$, error bars, s.e.m.). **c,** Fluorescent *in situ* hybridization for *Igf1* and *Vip* in mouse visual cortex after dark housing and light exposure for 7.5 h (white box indicates the magnified area; scale bar, 200 μm in main image; 50 μm in magnification). **d–f,** Quantification of fluorescent *in situ* hybridization for *Igf1* (**d**), *Crh* (**e**), *Prok2* (**f**) and inhibitory markers in visual cortices of dark-housed/light-exposed mice ($n = 3$, bars represent s.e.m.). GOI, gene of interest.

PV, SST, or pyramidal neurons (Extended Data Fig. 6e–l). Furthermore, ELISA-based analysis of IGF1 levels in the blood of mice whose cortices were injected with these viruses demonstrated that removing IGF1 from VIP neurons did not alter the level of serum-derived IGF1 (Extended Data Fig. 6d). While we formally cannot exclude that serum-derived IGF1 contributes to inhibition onto VIP neurons, this finding indicates that the decrease in mIPSCs in VIP neurons that express *Igf1* shRNAs is due at least in part to reduced *Igf1* expression in VIP neurons. Thus, VIP-neuron-derived IGF1 regulates the inhibitory inputs onto the VIP neuron in which it is produced, probably via local release from VIP neurons. Consistent with this idea, we find that the *Igf1* splice variant expressed by VIP neurons encodes an isoform of IGF1 containing a heparin binding domain (*Igf1.4*; Extended Data Fig. 7a) that may limit the diffusion of IGF1 and facilitate its local action[16].

We next overexpressed IGF1 in VIP neurons by injecting *Vip-cre* mice with an AAV construct that drives expression of an epitope-tagged version of IGF1 together with EGFP in a Cre-dependent manner (Fig. 3k and Extended Data Fig. 7b, c) and assessed the effect on mIPSCs and mEPSCs (Fig. 3l–o). We find that when overexpressed in VIP neurons, IGF1 selectively promotes inhibition onto VIP neurons, as it has no effect on mEPSCs in these cells. Likewise, ectopic expression of IGF1 in SST and excitatory neurons (by intracortical injections into *Sst*- or *Emx1-cre* mice, respectively; Extended Data Fig. 7d–g) leads to a similar increase in mIPSC frequency in these cells. These findings raise the possibility that the selective expression of *Igf1* in VIP neurons is required for the proper organization and function of cortical circuits, as aberrant IGF1 expression could enhance inhibition indiscriminately within cortical circuits by signalling through IGF1 receptors that are ubiquitously expressed in these neurons (Extended Data Fig. 7h–j).

The change in mIPSC frequency upon *Igf1* knockdown in VIP neurons could reflect a change in the presynaptic probability of release and/or a reduction in inhibitory synapse number and/or strength. By paired-pulse stimulation we find that *Igf1* knockdown in VIP neurons does not significantly alter the probability of release of vesicles from either inhibitory (Fig. 4a) or excitatory (Extended Data Fig. 7k) terminals that synapse onto VIP neurons. To test evoked inhibition, we co-injected *Vip-cre* mice with either *Igf1* or control shRNA AAVs and an AAV encoding the excitatory light-activated ion channel ReachR[17]. Performing paired voltage clamp recordings from eGFP-positive VIP neurons and neighbouring pyramidal cells to control for variation in stimulation intensity, we found that the strength of light-evoked inhibition onto VIP neurons is decreased when *Igf1* expression is knocked down in VIP neurons (Fig. 4b). These experiments suggest that the primary site of IGF1 action is post-synaptic and indicate that experience-dependent activation of *Igf1* expression increases the number and/or strength of functional inhibitory synapses that form on VIP neurons. To test whether this IGF1-dependent decrease in inhibition alters the frequency of action potentials in these neurons, we performed cell-attached recordings from eGFP-labelled VIP neurons expressing control or *Igf1* shRNAs. We find that VIP neurons lacking IGF1 fire action potentials at a significantly higher rate than controls (Fig. 4c). Given that VIP neurons disinhibit cortical circuits, it seems likely that this decreased firing of VIP neurons might alter how the cortex responds to sensory experience. To begin to investigate this possibility, we next assessed the effect of knocking down *Igf1* expression on visual cortex plasticity.

Cortical inhibition regulates ocular dominance (OD) plasticity[1,18] and visual acuity[19], and hyper-activation of VIP neurons drives a form of adult cortical plasticity[20]. To determine whether knocking down *Igf1* expression in VIP neurons affects visual cortex function, we injected control or *Igf1* shRNA AAVs into the binocular zone of visual cortices of P18 *Vip-cre* mice and recorded visual-evoked potentials between P28 and P32 (Fig. 4d and Extended Data Fig. 8a). Stimulation of the contralateral or ipsilateral eye with gratings at low spatial frequency elicited robust visual-evoked potentials both under control and *Igf1*-knockdown conditions (Extended Data Fig. 8b–d). Furthermore, the ratio between the contralateral and ipsilateral eye's response (C:I ratio) was similar in the presence or absence of *Igf1* (Extended Data Fig. 8e), indicating that basic visual cortex function is not obviously disrupted

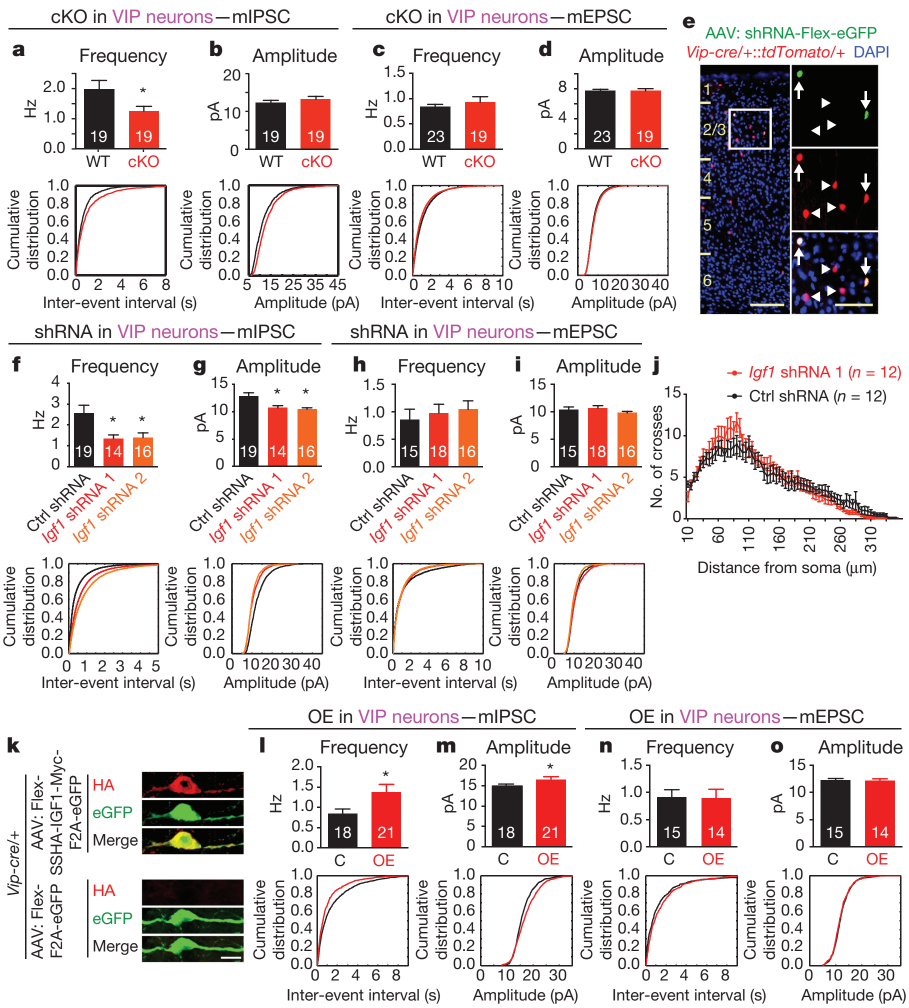


Figure 3 | IGF1 promotes inhibitory inputs to VIP neurons in a cell-autonomous manner. a–d, Bar graph and cumulative distribution of the frequency and inter-event intervals of mIPSCs or mEPSCs recorded from *Igf1* wild-type (WT) or conditional-knockout (cKO) VIP neurons (mIPSC frequency, $P = 0.046$; amplitude, $P = 0.3$; mEPSC frequency, $P = 0.44$; amplitude, $P = 0.9$, Mann–Whitney *U*-test). **e**, Example image of sparsely infected VIP neurons upon injection of AAV-shRNA-hUbc-Flex-eGFP into mice expressing tdTomato in all VIP neurons (white box indicates the magnified area; arrows indicate infected VIP neurons; arrowheads indicate non-infected VIP neurons; scale bars, 100 μm in main image; 50 μm in magnification). **f–i**, Bar graph and cumulative distribution of mIPSC/mEPSC frequency, inter-event interval and amplitude recorded from VIP neurons sparsely infected with control or *Igf1* shRNAs (mIPSC frequency: shRNA 1, $P = 0.05$; shRNA 2, $P = 0.042$; mIPSC amplitude: shRNA 1, $P = 0.004$; shRNA 2, $P = 0.001$; mEPSC frequency, shRNA 1, $P = 0.13$; shRNA 2, $P = 0.07$; mEPSC amplitude: shRNA 1, $P = 0.77$; shRNA 2, $P = 0.44$. Mann–Whitney *U*-test). **j**, Sholl analyses of VIP neurons infected with control or *Igf1* shRNA ($P = 0.76$, two-way repeated-measures ANOVA). **k**, Expression of epitope-tagged IGF1.4 in VIP-neurons. Cortices of P3 *Vip-cre*/+ mice were injected with AAVs driving Cre-dependent expression of SSHA-IGF1.4-Myc-F2A-eGFP (top) or F2A–eGFP (bottom) and stained at P20 for HA (red) and eGFP (green) (Scale bar, 10 μm). **l–o**, Bar graphs and cumulative distribution plots showing mIPSC/mEPSC amplitude and frequency/inter-event interval in VIP neurons infected with a control AAV or an AAV over-expressing (OE) IGF1. mIPSC: amplitude, $P = 0.05$; frequency, $P = 0.02$; mEPSC: amplitude, $P = 0.55$; frequency, $P = 0.86$. Mann–Whitney *U*-test. **a–d**, **f–j**, **l–o**, Numbers inside bars indicate the number of cells recorded; $*P < 0.05$ by Mann–Whitney *U*-test.

upon *Igf1* knockdown. Remarkably, when we assessed visual acuity of the contralateral eye by increasing the spatial frequency of the gratings presented, the mice injected with AAVs expressing *Igf1* shRNA exhibited significantly increased visual acuity as compared to mice injected with control AAVs (Fig. 4g).

To test whether the effect of *Igf1* knockdown is experience-dependent, we next monocularly deprived mice for a brief period of time, beginning at the peak time of ocular dominance plasticity (that is, at P26–28, Fig. 4d). After four days of monocular deprivation, we recorded visual-evoked potentials from the visual cortex contralateral to the deprived eye and quantified the C:I ratio upon stimulation at low spatial frequency as well as the visual acuity upon stimulation of the deprived eye. Brief monocular deprivation led to a reduction in the C:I ratio in mice injected with AAVs expressing either control or *Igf1* shRNA; this is a consequence of the reduction in the contralateral response (Extended Data Fig. 8b–e)[18]. Notably, when we tested visual acuity after brief monocular deprivation, both *Igf1* and control shRNA injected mice exhibited similar levels of amblyopia (that is, loss of visual acuity) in the deprived eye (Fig. 4e–g), despite the higher visual acuity in the *Igf1* shRNA injected mice that were not monocularly deprived (Fig. 4e–g). These findings indicate that VIP neuron-derived IGF1 regulates visual acuity in an experience-dependent manner and may function as a sensory-dependent brake on cortical plasticity. The observation that in response to sensory experience IGF1 in VIP neurons controls inhibition, taken together with the previous finding that experience induces BDNF in excitatory neurons to regulate excitatory–inhibitory balance[21,22], suggests a model in which each type of neuron within a cortical circuit expresses a unique set of experience-induced secreted factors that control specific synaptic inputs onto the neuron and plasticity within a neural circuit[6,23].

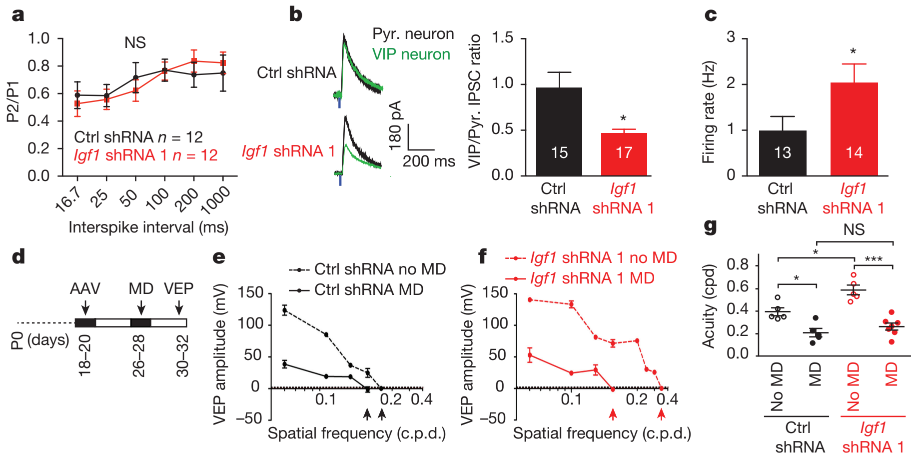


Figure 4 | VIP-neuron-derived IGF1 regulates VIP neuron function and regulates visual acuity in an experience-dependent manner. a, Paired-pulse recordings from VIP neurons infected with control or *Igf1* shRNA ($P = 0.96$, two-way ANOVA). **b**, Left, average traces of light-evoked IPSCs (eIPSC) from paired recordings of VIP neurons infected with control or *Igf1* shRNA (green traces) and neighbouring pyramidal neurons (Pyr., black traces). Right, quantification of eIPSC amplitude of the VIP neuron after infection with AAVs expressing control or *Igf1* shRNA, normalized to the eIPSC amplitude of the paired pyramidal neuron ($P = 0.01$, Mann–Whitney U-test). **c**, Average firing rate of VIP neurons infected with *Igf1* or control shRNA ($P = 0.04$, Mann–Whitney U-test). **b**, **c**, Numbers inside bars indicate the number of cells recorded. **d**, Schematic of the schedule for monocular deprivation (MD) experiments. **e**, **f**, Representative traces of visually evoked potential (VEP) amplitude as a function of spatial frequency (cycles per degree (c.p.d.)) in the contralateral visual cortex of mice that received bilateral injections of AAVs expressing *Igf1* or control shRNA into their visual cortices and were subjected to monocular deprivation or not. **g**, VIP-neuron-derived IGF1 restricts visual acuity in an experience-dependent manner. Visual acuity in mice injected with AAVs expressing *Igf1* or control shRNA with or without monocular deprivation (P24 – P28; control shRNA no MD, $n = 5$; control shRNA MD, $n = 5$; *Igf1* shRNA no MD, $n = 5$; *Igf1* shRNA MD, $n = 7$; $^{*}P < 0.05$; $^{***}P < 0.0001$; NS, not significant; one-way ANOVA with Tukey's post hoc test).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 June 2014; accepted 29 January 2016.

Published online 9 March 2016.

1. Hensch, T. K. Critical period plasticity in local cortical circuits. *Nature Rev. Neurosci.* **6,** 877–888 (2005).
2. Rudy, B., Fishell, G., Lee, S. & Hjerling-Leffler, J. Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. *Dev. Neurobiol.* **71,** 45–61 (2011).
3. Pfeffer, C. K., Xue, M., He, M., Huang, Z. J. & Scanziani, M. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature Neurosci.* **16,** 1068–1076 (2013).
4. Lee, S., Kruglikov, I., Huang, Z. J., Fishell, G. & Rudy, B. A disinhibitory circuit mediates motor integration in the somatosensory cortex. *Nature Neurosci.* **16,** 1662–1670 (2013).
5. Pi, H.-J. *et al.* Cortical interneurons that specialize in disinhibitory control. *Nature* **503,** 521–524 (2013).
6. Spiegel, I. *et al.* Npas4 regulates excitatory-inhibitory balance within neural circuits through cell-type-specific gene programs. *Cell* **157,** 1216–1229 (2014).
7. Majdan, M. & Shatz, C. J. Effects of visual experience on activity-dependent gene regulation in cortex. *Nature Neurosci.* **9,** 650–659 (2006).
8. Sanz, E. *et al.* Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proc. Natl Acad. Sci. USA* **106,** 13939–13944 (2009).
9. Kubota, Y. *et al.* Selective coexpression of multiple chemical markers defines discrete populations of neocortical GABAergic neurons. *Cereb. Cortex* **21,** 1803–1817 (2011).
10. Bondy, C. A. Transient IGF-I gene expression during the maturation of functionally related central projection neurons. *J. Neurosci.* **11,** 3442–3455 (1991).
11. Liu, J. P., Baker, J., Perkins, A. S., Robertson, E. J. & Efstratiadis, A. Mice carrying null mutations of the genes encoding insulin-like growth factor I (Igf-1) and type 1 IGF receptor (Igf1r). *Cell* **75,** 59–72 (1993).
12. Cheng, C. M. *et al.* Insulin-like growth factor 1 is essential for normal dendritic growth. *J. Neurosci. Res.* **73,** 1–9 (2003).
13. Cao, P., Maximov, A. & Südhof, T. C. Activity-dependent IGF-1 exocytosis is controlled by the Ca^{2+}-sensor synaptotagmin-10. *Cell* **145,** 300–311 (2011).
14. Nishijima, T. *et al.* Neuronal activity drives localized blood-brain-barrier transport of serum insulin-like growth factor-I into the CNS. *Neuron* **67,** 834–846 (2010).
15. Liu, J. L. *et al.* Insulin-like growth factor-I affects perinatal lethality and postnatal development in a gene dosage-dependent manner: manipulation using the Cre/*loxP* system in transgenic mice. *Mol. Endocrinol.* **12,** 1452–1462 (1998).
16. Hede, M. S. *et al.* E-peptides control bioavailability of IGF-1. *PLoS ONE* **7,** e51152 (2012).
17. Lin, J. Y., Knutsen, P. M., Muller, A., Kleinfeld, D. & Tsien, R. Y. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nature Neurosci.* **16,** 1499–1508 (2013).
18. Fagiolini, M. & Hensch, T. K. Inhibitory threshold for critical-period activation in primary visual cortex. *Nature* **404,** 183–186 (2000).
19. Davis, M. F. *et al.* Inhibitory neuron transplantation into adult visual cortex creates a new critical period that rescues impaired vision. *Neuron* **86,** 1055–1066 (2015).
20. Fu, Y., Kaneko, M., Tang, Y., Alvarez-Buylla, A. & Stryker, M. P. A cortical disinhibitory circuit for enhancing adult plasticity. *eLife* **4,** e05558 (2015).
21. Hong, E. J., McCord, A. E. & Greenberg, M. E. A biological function for the neuronal activity-dependent component of Bdnf transcription in the development of cortical inhibition. *Neuron* **60,** 610–624 (2008).
22. Bloodgood, B. L., Sharma, N., Browne, H. A., Trepman, A. Z. & Greenberg, M. E. The activity-dependent transcription factor NPAS4 regulates domain-specific inhibition. *Nature* **503,** 121–125 (2013).
23. Turrigiano, G. Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Annu. Rev. Neurosci.* **34,** 89–103 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Chen for help with electrophysiology experiments, E. Griffith and T. Cherry for critical reading of the manuscript, P. Zhang for managing the mouse colony and the HMS Biopolymers Facility Next-Gen Sequencing Core for their expertise in constructing Seq libraries and sequencing of the library samples. The ReachR-tdTomato construct was a gift from J. Lin, and we thank M. Li for production of the ReachR-virus. H.A. is a New York Stem Cell Robertson Investigator. This work was funded by fellowships by the Human Frontiers Science Program and the Swiss National Science Foundation (I.S.) and the National Institute of Health grants R01 NS028829 and P01 NS047572 (M.E.G.).

Author Contributions Experiments were designed by A.R.M., I.S. and M.E.G. Experiments were conducted and analysed by A.R.M., I.S., E.C., A.P., C.P.T., J.E.B., C.M.B. and D.A.H. Experiments were supervised by H.A., M.F. and M.E.G. The manuscript was prepared by A.R.M., I.S. and M.E.G.

Author Information Raw data and processed values from RiboTag-Seq have been submitted to the NCBI Gene Expression Omnibus under the accession number GSE77243. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.E.G. (Michael_Greenberg@hms.harvard.edu).

METHODS

No statistical methods were used to predetermine sample size.

Visual stimulation. For calibrating the duration of the dark housing period before light exposure, C57Bl6 wild-type mice were housed in a standard light cycle until they were placed in constant darkness for varying amounts of time before analysis at postnatal day 56. At P56, all mice were either sacrificed in the dark (dark-housed condition) or light-exposed for 1, 3, or 7.5 h before being sacrificed. The eyes of all animals were enucleated (for the dark-housed condition, enucleation was performed in the dark) before dissection of the visual cortex in the light.

For RiboTag-experiments, mice were reared in a standard light cycle and then housed in constant darkness for two weeks starting from P42; at P56, all mice were either sacrificed in the dark (dark-housed condition) or light-exposed for 1, 3, or 7.5 h before being sacrificed. Additional cohorts of mice for the 'standard' condition were housed in a standard light cycle until P56 when they were euthanized. The eyes of all animals were enucleated (for the dark-housed condition, enucleation was performed in the dark) before dissection of the visual cortex in the light.

RNA isolation, reverse transcription, qPCR analysis. Total RNA was extracted with TRIzol reagent (Sigma) according to the manufacturer's instructions, and RNA quality was assessed on a 2100 BioAnalyzer (Agilent); all RNAs were treated with DNaseI (Invitrogen) before reverse transcription. For the cloning of riboprobes, total RNA was extracted from whole adult C57Bl6 wild-type mouse brains and cDNA was prepared using SuperScript II kit (Life Technologies). For real-time quantitative PCR experiments aimed at calibrating the duration of the dark housing period, total RNA was extracted for each sample from the visual cortices of one animal. For real-time quantitative PCR experiments aimed at testing the efficacy of shRNA constructs directed against *Igf1*, total RNA was isolated from two pooled 24 wells of cultured cortical neurons for each condition. For qPCR experiments, RNA was reverse-transcribed with the High Capacity cDNA Reverse Transcription kit (Life Technologies). Real-time quantitative PCR reactions were performed on the LightCycler 480 system (Roche) with LightCycler 480 SYBR Green I Master. Reactions were run in duplicates, triplicates or quadruplicates, and β-actin (*Actb*) or β3-tubulin (*Tubb3*) levels were used as an endogenous control for normalization using the $\Delta\Delta C_t$ method[24]. Real-time PCR primers were designed using the Universal ProbeLibrary (Roche) as exon-spanning whenever possible and answered the following criteria: linear amplification over three orders of magnitude of target concentration, no amplification product in control samples that were not reverse-transcribed (that is, control for contamination with genomic DNA), no amplification product in control samples where no template was added (that is, control for primer dimers), amplification of one singular product as determined by melt-curve analysis and analysis of the product in agarose gel electrophoresis and sequencing of the PCR product. The qPCR primers used in this study are listed in Supplementary Table 6.

For analysis of light-induced gene expression in wild-type mice, the gene expression levels were analysed in four mice (two males and two females) at each time point. The data were calculated as fold change relative to the average of the overnight dark-housed condition and normalized to the average of the maximally induced time point. Data in figures represent the mean and s.e.m. of four mice.

For assessing *Igf1* levels in cortical cultures infected with shRNA-expressing lentiviral constructs, qPCRs were performed in quadruplicates for each condition and fold changes were calculated relative to the non-infected non-stimulated cultures. Data were normalized to the maximally induced condition in each biological replicate, and data in figures represent the mean and s.e.m. of three biological replicates.

RiboTag-purifications, RiboTag-qPCR and RiboTag-seq. Immunoprecipitation and purification of ribosome associated RNA was performed essentially as described[6,8], with minor modifications: lysis of the samples was performed in the presence 10 mM Ribonucleoside Vanadyl Complex (NEB, Ipswich, MA), and immunoprecipitation was performed with a different anti-HA antibody (HA-7, 12 μg per immunoprecipitation, Sigma). In brief, the visual cortices were dissected, flash frozen in liquid nitrogen and then kept at −80 °C until further processing. Visual cortices from three individual animals (each sample contained both male and female animals) were pooled for each biological replicate, and three biological replicates were performed. After lysis of the tissues and before immunopurification, a small fraction of lysate of each sample (that is, 'input') was set aside and total RNA was extracted with TRIzol reagent followed by the RNEasy Micro Kit's procedure (Qiagen, Valencia, California). After immunopurification of the ribosome-associated RNAs, RNA quality was assessed on a 2100 BioAnalyzer (Agilent, Palo Alto, California) and RNA amounts were quantified using the Qubit 2.0 Fluorometer (Life Technologies). Only samples with RIN numbers above 8.0 were considered for analysis by qPCR and RNA-seq. For all RNA samples of sufficient integrity, 5–10 ng of RNA were SPIA-amplified with the Ovation RNA Amplification System V2 (NuGEN, San Carlos, California), yielding typically 5–8 μg of cDNA per sample.

Quantitative RT–PCR was performed as described above and relative expression levels were determined in every experiment by normalizing the Ct-values to those of beta-Actin (*ActB*) from the 0 h input using the $\Delta\Delta C_t$ method[24]. To determine the fold-enrichment (IP/Input), the actin-normalized expression levels for every time point of every biological replicate were averaged, and the grand averages from the IP and Input were divided to find the IP/Input ratio. To calculate fold-induction for each biological replicate, each time point was divided by the maximal value occurring in that biological replicate, such that the maximal value was set to 1 in each biological replicate. The mean and standard error were calculated at each time point from these normalized values. All samples were analysed by qPCR for purity and light-induced gene expression before analysis by high throughput sequencing.

RNA-seq and analysis. SPIA-amplified samples from RiboTag-immunoprecipitated fractions for each of the five stimulus conditions and each of the five Cre lines were prepared as described above and processed in triplicate (75 samples total). For preparing sequencing libraries, 2 μg of each amplified cDNA were fragmented to a length of 200–400 bp using a Covaris S2 sonicator (Acoustic Wave Instruments) using the following parameters: duty cycle: 10%, intensity: 5, cycles per burst: 200, time: 60 s, total time: 5 min. After validating the fragment length of the sonicated cDNA using a 2100 BioAnalyzer (Agilent, Palo Alto, California), 2 μg of the fragmented cDNA were used for sequencing library preparation using the PrepX DNA kit on an Apollo 324 robot (IntegenX). The quality of completed sequencing libraries was assessed using a 2100 BioAnalyzer (Agilent, Palo Alto, California) and the completed libraries were sequenced on an Illumina HiSeq 2000 instrument, following the manufacturer's standard protocols for single-end 50 bp sequencing with single index reads. Sequencing typically yielded 30–80 million usable non-strand-specific reads per IP sample. Reads were mapped to the mm9 genome using TopHat (v.2.0.13) and Bowtie (2.1.0.0)[25]. On average, ~70% of mapped IP reads were uniquely mapped to the mm9 genome allowing for 0 mismatches and were therefore assignable to genic features (one RiboTag-seq library (*Sst-cre*, standard-housing, biological replicate 2) was excluded from analysis due to low mappability). Values from all IP libraries were normalized using Cufform (v.2.2.1), and values from the Cuffnorm output file 'genes-Count_Table' (normalized reads) were taken as a proxy for gene expression. *P* values were generated for each Cre line for each dark–light conditions using Cuffdiff (v.2.2.1) using the time series (-T) flag based on three biological replicates.

Identification and classification of experience-regulated transcripts. To identify transcripts regulated by visual experience, for each biological replicate of each Cre line, the fold change in normalized reads was calculated for each gene at every time point (dark-housed/standard-housed, 1 h light/dark-housed, 3 h light/dark-housed; 7.5 h light/dark-housed). Genes were flagged as experience-regulated in a given Cre line if they met the following conditions in at least one sample: (1) P value <0.005, (2) mean fold change of twofold or greater, (3) fold changes of 2 or higher in 2 of 3 biological replicates, (4) the mean expression value in at least one sample must be above absolute expression threshold (set at the 40th percentile of all observed values).

To determine in which Cre lines genes were regulated by experience, genes were simply classified according to the above criteria. However, for this analysis we excluded the *Gad2-cre* line, since *Pv-*, *Sst-* and *Vip-cre* all label subsets of the neurons labelled by *Gad2-cre*. However, we did detect genes regulated solely in *Gad2-cre*, but no other Cre lines; we reasoned that these genes are probably regulated by experience in a population of $5HT3aR^+/VIP^-$ neurons that are contained in *Gad2-cre* but none of the other Cre lines.

We classified the set of experience-regulated genes into categories 'early', 'late', and 'long-term' based on the fastest kinetics observed. When genes were found to be elevated and/or suppressed at multiple time points, we assigned them to the categories based on the most rapid observed change. For example, while *Fos* levels are elevated over dark housing at 1, 3 and 7.5 h of light exposure and suppressed after two weeks of dark housing, *Fos* is classified as 'early-up' because it is elevated at 1 h after light exposure.

Linkage analysis. All linkage analysis was performed using the 'single' method and 'Cityblock' metric using Matlab's linkage function. To determine the branch-order significance of the cladogram resulting from clustering of the 602 experience-regulated genes, we generated 1,000 cladograms from 602 sets of random expressed genes (including experience-regulated genes, with replacement) and asked how often we generated a cladogram with an identical branch order at the level of the Cre lines. Only 11 sets of 1,000 random genes sets generated an identical tree. For the purposes of this analysis, we only compared the branches above the level of the individual Cre line.

Identification of cell-type-specific transcripts. To identify cell-type-enriched transcripts, an enrichment score was calculated for every transcript in every Cre line for each biological replicate. This enrichment score was calculated by dividing the maximum expression value observed in a given Cre line by the maximum expression value observed across all conditions for all other Cre lines

(GABAergic subtypes were not required to be enriched above *Gad2-cre*). The enrichment scores for a set of known cell-type-specific genes were evaluated (*Vglut1*, *Tbr1*, *Pvalb*, *Sst*, *Vip*), and our threshold was set at the enrichment score of the cell-type-specific gene with the lowest score (*Slc17a7/Vglut1*, at 5.5-fold-enriched in *Emx1-cre*). Transcripts were considered to be expressed in a cell-type-specific manner (or 'highly enriched') in a given Cre line if their mean enrichment score was above this threshold and if the enrichment score exceeded this threshold in 2 out of 3 biological replicates.

Cloning of riboprobes, knockdown and expression constructs. Cloning of all constructs was done using standard cloning techniques, and the integrity of all cloned constructs was validated by DNA sequencing. Templates for the riboprobes for *Igf1*, *Gad1*, *Pvalb*, *Sst* and *Vip* were prepared by PCR-amplification of cDNA fragments generated from total RNA isolated from adult C57Bl6 mouse brains (see Supplementary Table 7 for primer sequences) and cloning of the respective PCR fragments into the pBlueScript II vector (Agilent Technologies).

Lentiviral shRNA constructs were generated by cloning shRNA stem loop sequences against *Igf1* (*Igf1* shRNA 1: GGTGGATGCTCTTCAGTTC; *Igf1* shRNA 2: TGAGGAGACTGGAGATGTA) and Luciferase (*Luc*, control: ACTTACGCTGAGTACTTCG) into a modified version of pLentiLox3.7[26] in which the CMV promoter driving the expression of eGFP was replaced with an hUbc promoter and in which the *loxP* sites surrounding the hUbc-eGFP cassette were removed. The loop sequence used in these shRNA constructs is based on miR-25 (CCTCTCAACACTGG)[27]. shRNA-expressing AAV-constructs (pAAV-U6-shRNA-hUbc-Flex-eGFP) were made by first replacing the Flex-GFP-Gephyrin cassette in pAAV-Flex-GFP-Gephyrin[22] with a Flex-eGFP cassette (resulting in pAAV-hUbc-Flex-eGFP) and then transferring the U6-shRNA cassettes from the pLentiLox constructs to pAAV-hUbc-Flex-eGFP.

AAV constructs for the Cre-conditional co-expression of epitope-tagged IGF1.4 and eGFP or of eGFP alone were cloned by synthesizing the gBlocks (Integrated DNA Technologies) and using the gBlocks as templates for PCR amplification; the respective PCR products were then cloned into the pAAV-hUbc-Flex-eGFP (see above) by replacing the EGFP with the respective insert. This strategy yielded plasmids termed pAAV-hUbc-Flex-SSHA-IGF1.4-Myc-F2A-eGFP and pAAV-hUbc-Flex-F2A-eGFP, whereby the Cre-dependent inserts were driven by a human ubiquitin promoter (hUbc). The sequence for *Igf1.4* was based on NM_001111275 (base pairs 277–752) and was modified in the following way: an HA epitope (TATCCtTATGATGTTCCAGATTATGCT) was inserted in frame between the *Igf1.4* signal sequence and the beginning of the coding sequencing (cds) of *Igf1.4*, *Igf1.4* was rendered resistant to the shRNA against *Igf1* by introducing silent mutations into the target sequences specified above (sh1: TGTTGACGCGCTCCAATTT; sh2: TACGCCGGTTAGAAATGTA) and the followings tags were inserted in frame 3′ to the *Igf1.4* coding sequencing: Myc epitope (GAACAAAAACTCATCTCAGAAGAGGATCTG), Furin cleavage site (CGGGCCAAGCGG) and a 2A peptide (GGCAGTGGAGAGGGCAGAGGA AGTCTTCTAACATGCGGTGACGTGGAGGAGAATCCCGGCCCT). The sequence for eGFP was based of the published sequence of eGFP. For pAAV-hUbc-Flex-F2A-eGFP a gBlock was synthesized containing the Furin cleavage site followed by the 2A site and eGFP. Detailed sequences are available upon request.

Double-fluorescent ISH. For double-fluorescent *in situ* hybridization (FISH), wild-type C57Bl6 mice were dark-housed and light-exposed for 7.5 h as described above. After light exposure, the brains were dissected and fresh frozen in Tissue-Tek Cryo-OCT compound (Fisher Scientific) on dry ice and stored at −80 °C until use.

FISH for *Igf1* was essentially done as described[28,29]: riboprobes were prepared by *in vitro* transcription of linearized plasmids containing the template of the respective probe. Riboprobes for *Igf1* were labelled with UTP-11-Digoxigenin, while the riboprobes for the subtype markers (*Gad1*, *Pvalb*, *Sst*, *Vip*) were labelled with UTP-12-Fluorescein (Roche); all riboprobes were hydrolyzed to lengths of 200–400 bp after synthesis and validated for labelling with Dioxigenin or Fluorescein. For *in situ* hybridization, coronal sections (20 μm thick) of the visual cortex were cut on a cryostat and fixed in 4% paraformaldehyde for 10 min. Endogenous peroxidases were inactivated by treating the sections for 15 min in 0.3% H_2O_2 in PBS, and acetylation was performed as described. Pre-hybridization was done overnight at room temperature, and hybridization was performed under stringent conditions at 71.5 °C. Following hybridization, stringency washes in SSC were performed as described at 65 °C. For immunological detection of the first probe (*Igf1*), the tissue was first treated with a blocking step for 1 h in blocking buffer (B2) at room temperature before the anti-Digoxigenin-POD antibody (Roche) was applied at a concentration of 1:1000 in blocking buffer for 1 h at room temperature. Following three washes in buffer B1 and an additional wash in buffer TNT (0.1 M Tris-HCl pH 7.5, 0.15 M NaCl, 0.05% Tween20), the *Igf1* probe was detected by exposing the sections at room temperature in the dark for 20 min to TSA Plus Cy3 reagent (Perkin Elmer) diluted 1:100 in TSA working solution, after which the sections were washed three times in TNT buffer. Before the immunological detection of the second probe, the peroxidases for detecting the first probe were inactivated by treating the sections for 30 min with 3% H_2O_2, followed by three washes in PBS. After an additional blocking step in blocking buffer for 1 h at room temperature, the anti-fluorescein-POD antibody (Roche) was applied at a concentration of 1:1000 in blocking buffer overnight at 4 °C. Following three washes in buffer B1 and an additional wash in buffer TNT, the probes of the subtype markers were detected by exposing the sections at room temperature in the dark for 15 min to TSA Plus Cy5 reagent (Perkin Elmer) diluted 1:100 in TSA working solution, after which the sections were washed three times in TNT buffer. Finally, the sections were counterstained with DAPI (4′,6-diamidino-2-phenylindole, Molecular Probes) and mounted using Fluoromount-G (Southern Biotech). In each experiment, controls for hybridization specificity were included (sense probe for *Igf1*) as well as controls for ensuring the specificity of the immunological detection of the digoxigenin- and fluorescein-labelled riboprobes.

FISH for *Crh*, *Prok2* and *Fbln2* was done using the RNAscope system (Advanced Cell Diagnostic); this was necessary since no reliable signal could be detected with the method described above for *Igf1* FISH using DIG-labelled riboprobes. RNAscope probes for all genes were synthesized by ACD and all experiments were done according to the ACD's protocol for fresh frozen brain sections.

For quantifying of the expression pattern of all genes of interest (GOI, that is, *Igf1*, *Crh* and *Prok2*; *Fbln2* could not be detected reliably), the visual cortices in each section were imaged on a Zeiss Axio Imager microscope with a 10× objective and 3 × 5 fields-of-view were 'stitched' into one compound image; in all cases, image exposures were kept constant throughout a given experiment for each channel. Compound images of each visual cortex were then imported to Photoshop, and additional layers were created for each probe (that is, one layer for the GOI and one layer for the subtype marker in each compound image). The cells positive for each probe were then marked with a dot in the new respective layer by two independent investigators in a blinded manner (one investigator marking GOI-positive cells and the other investigator marking subtype-marker-positive cells). Finally, the layers containing the dots of the identified positive cells were compiled into a separate image file together with the DAPI-layer and imported into ImageJ. In ImageJ, the images were analysed in a blinded manner by defining the visual cortex and its layers as regions of interest (ROI) based on the DAPI staining and quantifying the number of cells positive for either one or both markers per ROI. For each combination of probes (GOI together with each of the subtype markers), two visual cortices from four animals were analysed (a total of eight visual cortices for each combination).

Virus production and neuronal cultures. Concentrated lentiviral stocks were prepared and titrated essentially as described[30]. AAV stocks were prepared at the University of North Carolina (UNC) Vector Core and at the Children's Hospital Boston Vector Core; see also Supplementary Table 8 for further details on AAV stocks.

Primary cultures of cortical neurons were prepared from E16.5 mouse embryos as described[6]. In brief, 3×10^5 neurons per well were plated in 24-well dishes coated with poly-D-lysine (20 μg ml^{-1}) and laminin (3.4 μg ml^{-1}). Cultures were maintained in neurobasal medium supplemented with B27 (Invitrogen), 1 mM L-glutamine, and 100 U ml^{-1} penicillin/streptomycin, and one-third of the media in each well was replaced every other day. For testing of viral shRNA constructs, the cultures were infected at DIV 3 with concentrated viral stocks for 5 h at an MOI of 6. After infection, the cultures were washed twice in plain neurobasal medium after which the conditioned medium was returned to the dish and the cultures were continued to be maintained as described. At DIV 7, neuronal cultures were treated overnight with 1 μM TTX and 100 μM AP-5 to silence spontaneous activity before the cultures were depolarized at DIV 8 with 55 mM extracellular KCl as described[6] and lysed in TRIzol after 6 h of stimulation.

Western blot for testing of IGF1 constructs and ELISA for determining serum IGF1 levels. HEK293T cells were used for testing the expression and the biological activity of the epitope-tagged IGF1.4 constructs. HEK293T cells were cultured in DMEM (Life Sciences) containing 10% FCS and penicillin/streptomycin. Cells were transfected using lipofectamine (Life Technologies) and 18 h post transfection, the medium was replaced with DMEM containing 0.1% FCS; 42 h post transfection, the conditioned media were collected, spun down to remove cell debris and used immediately for stimulating non-transfected HEK293T that were serum starved for 3 days in DMEM containing 0.1% FCS. The conditioned media were applied to the serum starved cells for 15 min at 37 °C after which the cells were lysed in boiling SDS sample buffer and subjected to Western blot analysis essentially as described[6,31]. For detecting the (phosphorylated) IGF1-receptor, the following antibodies were used: anti-IGF1-receptor-β (D23H3) XP Rabbit mAb (#9750, Cell Signaling, 1:1000) and anti-phospho-IGF1-receptor-β (Tyr1135/1136)/Insulin Receptor β (Tyr1150/1151) (19H7) Rabbit mAb (#3024, Cell Signaling, 1:1000). For determining serum IGF1 levels, we used the IGF1 Quantikine ELISA kit (R&D

Systems), following the manufacturer's instructions (P3 *Vip-cre* heterozygous pups were injected intracortically with the respective AAV and bled at P20).

Perfusions, immunohistochemistry and morphological analysis of IGF1 cKO visual cortices. Mice were anaesthetized with 10% ketamine and 1% xylazine in PBS by intraperitoneal injection. When fully anaesthetized, the animals were transcardially perfused with ice-cold PBS for 5 minutes followed by 15 minutes of cold 4% PFA in PBS. Brains were dissected and post-fixed for one hour at 4 °C in 4% PFA, followed by three washes (each for 30 min) in cold PBS, and cryoprotection overnight in 20% sucrose in PBS at 4 °C. The following day, brains were placed in Tissue-Tek Cryo-OCT compound (Fisher Scientific), frozen on dry ice and stored at −80 °C. Coronal sections (20 μm thick) of the visual cortices were subsequently cut using a Leica CM1950 cryostat and used for subsequent experiments.

For immunolabelling, the slides were blocked for 1 h with PBS containing 5% normal goat serum and 0.1% Triton X-100 (blocking solution). The samples were incubated overnight with different primary antibodies diluted in blocking solution, washed three times with PBS and then incubated for 45 min at room temperature with secondary antibodies and/or Hoechst stain (ThermoFisher Scientific). Slides were mounted in FluoromountG (Southern Biotech) and imaged on a Zeiss Axio Imager microscope. The following antibodies were used: mouse anti-HA (HA-7, Sigma; 1:1000), chicken anti-GFP (GFP-1020, Aves labs ; 1:1500), goat anti-mouse IgG (H+L) Alexa Fluor 488 (Highly Cross-Adsorbed, Life Technologies; 1:1,000), goat anti-chicken IgY (H+L) Alexa Fluor 488 (Life Technologies; 1:1,000).

For analysing the brains of *Igf1 Vip-cre* WT and cKO mice, brains of three-week-old WT and cKO littermates were placed on the same slide to minimize variation. After cryosectioning, the slides were either counterstained immediately or stored at −20 °C before they were counterstained and imaged. Counterstaining was done with DAPI (4′,6-diamidino-2-phenylindole, Molecular Probes) in PBS for 15–30 min at room temperature, after which the sections were washed once in PBS and mounted in Fluoromount-G (SouthernBiotech). For cell counting experiments, coronal visual cortex sections were imaged using a Zeiss Axio Imager microscope with a 10× objective and typically, 3 × 5 fields-of-view were 'stitched' into one compound image. In all cases, image exposures were kept constant throughout a given experiment for each channel. Custom ImageJ and MATLAB macros were used to quantify the area of each cortical layer, the number tdTomato-positive cells per layer, and the size of tdTomato-postive cells. Briefly, regions of interest (ROI) encompassing the visual cortex and its layers were defined based on the DAPI counterstaining. While the width of these ROIs was kept constant throughout the analysis of all sections, the height of the ROIs was adjusted in each image according to the DAPI counterstaining in each section and the areas of each layer in each section were recorded. For analysing the number and soma size of tdTomato-postive cells in each layer, a threshold for each channel was determined based on multiple user-defined negative regions. Channels were thresholded and binarized, and a mask of each channel was created. The number of tdTomato-positive cells was determined by taking the logical AND of the DAPI and tdTomato channel masks and counting the number of components greater than 4 pixels in size in the double overlap of the masks of the two channels in each layer ROI. The soma size was calculated as the area of these double-overlapping components. Three animals per genotype and 4–6 visual cortex sections per animal were analysed, and these data were used to determine the mean and s.e.m. of the values reported for each genotype.

Stereotaxically guided surgery and intra-cortical injections of AAV constructs. All surgeries were performed according to protocols approved by the Harvard University Standing Committee on Animal Care and were in accordance with federal guidelines. Surgeries were performed on mice between P14 and P15. Animals were deeply anaesthetized by inhalation of isoflurane (initially 3–5% in O_2, maintained with 1–2%) and secured in the stereotaxic apparatus (Kopf). Animal temperature was maintained at 37 °C. The fur was shaved and scalp cleaned with betadine and 100% ethanol three times before an incision was made to expose the skull. Injections into the visual cortex were made by drilling a ~0.5 mm burr hole (approximately 2.7 mm lateral, 0.5 mm anterior to lambda) through the skull, inserting a glass pipette to a depth of 200–400 μm and injecting 250 nl of the respective AAV construct at a rate of 100 nl min^{-1}. Five minutes post-injection, the glass pipette was retracted, the scalp sutured and the mouse returned to its home cage. All animals were monitored for at least one hour post-surgery and at 12 h intervals for the next 5 days. Post-operatively, analgesic (flunixin, 2.5 mg per kg) was administered at 12 h intervals for 72 h.

For neonatal injections, pups post-natal day 3–5 were anaesthetized on ice for 2–3 min, and secured to a stage where their head was supported using a clay mould using standard lab tape. A bevelled glass pipette was lowered into visual cortex (approximately 2 mm lateral, 0.2 mm anterior to lambda), and 50 nl of the respective AAV virus was injected at a rate of 23 nl sec^{-1}. Injections were made into eight sites (four on each hemisphere), and the mouse was then allowed to recover on a 37 °C warm plate before being returned to the home cage.

For bilateral stereotaxic intra-cortical injections of AAV constructs for visual plasticity experiments, surgeries were performed on mice between P18 and P20. Animals were anaesthetized with isofluorane gas (1–2% in O_2), and body temperature was maintained at around 37 °C with a heating pad during surgery. The head was held in place by standard mouse stereotaxic frame. The fur was shaved and scalp cleaned with betadine and 100% ethanol three times before an incision was made to expose the skull. Burr holes were drilled into the skull at the point of injection guided by stereotaxic coordinates and blood vessel patterns (approximately 2 mm and 2.7 mm lateral, 0.5 mm anterior to lambda) on both hemispheres. A 28-gauge Hamilton syringe (701RN) was inserted to a depth of 200–300 μm and 250 nl of the respective AAV construct was injected at the rate of 50 ml min^{-1}. Five minutes post-injection, the Hamilton syringe was retracted, the scalp sutured and the mouse returned to its home cage. All animals were monitored for at least one hour post-surgery. Post-operatively, analgesic (meloxicam, 5–10 mg kg^{-1}) was administered every 24 h for 2 days.

Electrophysiology. Coronal sections (300 μm thick) containing the primary visual cortex were cut from P19-P21 mice using a Leica VT1000S vibratome in ice-cold choline dissection media (25 mM $NaHCO_3$, 1.25 mM NaH_2PO_4, 2.5 mM KCl, 7 mM $MgCl_2$, 25 mM glucose, 0.5 mM $CaCl_2$, 110 mM choline chloride, 11.6 mM ascorbic acid, 3.1 mM pyruvic acid). Slices were incubated in artificial cerebral spinal fluid (ACSF, contains 127 mM NaCl, 25 mM $NaHCO_3$, 1.25 mM NaH_2PO_4, 2.5 mM KCl, 2.5 mM $CaCl_2$, 1 mM $MgCl_2$, 25 mM glucose) at 32 °C for 30 min immediately after cutting, and subsequently at room temperature. All solutions were saturated with 95% O_2/5% CO_2, and slices were used within 6 h of preparation. Whole-cell voltage-clamp recordings were performed in ACSF at room temperature from neurons in primary visual cortex that were identified under fluorescent and DIC optics. Recording pipettes were pulled from borosilicate glass capillary tubing with filaments using a P-1000 micropipette puller (Sutter Instruments) and yielded tips of 2–5.5 MΩ resistance. All experiments were recorded with pipettes filled with 135 mM caesium methanesulfonate, 15 mM HEPES, 0.5 mM EGTA, 5 mM TEA-Cl, 1 mM $MgCl_2$, 0.16 mM $CaCl_2$, 2 mM Mg-ATP, 0.3 mM Na-GTP, 10 mM phosphocreatine (Tris), and 2 mM QX-314-Cl. Osmolarity and pH were adjusted to 310 mOsm and 7.3 with Millipore water and CsOH, respectively. Recordings were sampled at 20 kHz and filtered at 2 kHz. mEPSCs were isolated by holding neurons at −70 mV and exposing them to 0.5 μM tetrodotoxin, 50 μM picrotoxin and 25 μM cyclothiazide and were blocked by application of 25 μM NBQX and 50 μM CPP. mIPSCs were isolated by holding neurons at 0 mV and exposing them to 0.5 μM tetrodotoxin, 25 μM NBQX, and 50 μM CPP and were blocked by 50 μM picrotoxin. Data were acquired using either Clampex10 or custom MATLAB software, using either an Axopatch 200B or Multiclamp 700B amplifier, and digitized with a DigiData 1440 data acquisition board (Axon Instruments) or a PCIe-6323 (National Instruments). For measuring miniature postsynaptic currents (minis), cells were allowed to stabilize for at least two minutes.

For paired pulse experiments, no drugs were used in the ACSF. A stimulating electrode (ISO-Flex, A.M.P.I.) was positioned approximately 100 μm below the cell, and 0.1 ms electrical pulses were given while adjusting the stimulus intensity and electrode position until the first pulse was between 100 and 500 pA. Inter-stimulus interval was varied and 10 s elapsed between each sweep. Pulse amplitudes were obtained from average sweeps of at least ten trials. Cells were held at 0 mV to record IPSCs and −70 mV to record EPSCs.

For evoked IPSCs, no drugs were used in the ACSF. Simultaneous paired whole-cell recordings were obtained from an eGFP-expressing VIP neuron and a morphologically identified pyramidal neuron located not more than five cell bodies away from the VIP neuron. Both cells were held at 0 mV, and a 5 ms light pulse from a blue LED (Thorlabs) was used to evoke IPSCs. Light intensity and the objective position were varied until the VIP neuron IPSC amplitude was between 200 and 500 pA. Average light power at 470 nm varied from between 0.3 and 0.7 mW over the course of the experiment. Reported ratios were obtained by dividing IPSC amplitudes obtained from an average trace of at least ten trials.

For electrophysiology experiments, n was set to min $n = 10$ to detect 20% effect size with power 0.95. For experiments to determine average firing rate of VIP neurons, a modified ACSF that promotes increased action potential firing was used containing, 3.5 mM KCl and 0.8 mM $CaCl_2$. Cell-attached patch recordings were obtained from eGFP-positive cells. Cells that did not fire an action potential in the first 30 s of recording were discarded, and recordings were maintained for at least 30 ten-second sweeps. Average firing rate was determined from the first sweep to the last recorded sweep in which an action potential occurred.

Miniature IPSC and EPSC data were analysed using Axograph X. Events were identified using a variable amplitude template-based strategy. Templates for each event type were defined as follows: mEPSC: 0.25 ms rise time, 3 ms decay τ, amplitude threshold of $-3 \times$ s.d. local noise; mIPSC: 1 ms rise time, 50 ms decay τ, amplitude threshold of $2.5 \times$ s.d. local noise. Local noise was determined by calculating the standard deviation of the current in a 5 ms window before event rise onset.

Templates lengths extended 25 ms after rise onset in the case of mEPSCs and 50 ms after rise onset in the case of mIPSCs. Events were discarded if they had a rise time outside the range of 0–3 ms. Statistical significance for all recorded parameters between genotypes was evaluated using a Mann–Whitney *U*-test on the mean values from individual neurons in a given experiment.

Minis were additionally evaluated for significance using both a Kolmogorov–Smirnov test (KS test) and Monte Carlo test. For these tests, 50 random minis were sampled from each neuron in each condition to obtain a continuous distribution for each condition that equally weighted each cell in that condition: these distributions are the data shown in the cumulative distribution graphs. One hundred random events were randomly sampled from these distributions for a KS test; and for Monte Carlo tests, 100 random events were randomly sampled from each distribution 1,000 times (with replacement), and the means were compared. All significant differences in mini amplitude and frequency were found to be significant by Monte Carlo test, KS test, and Mann–Whitney *U*-test of cell means. Since the Mann–Whitney test was found to be the most stringent test, the *P* values from Mann–Whitney tests are reported. All data was analysed blind to genotype or experimental condition. In all conditions, series resistance, holding potential, cell capacitance, and input resistance were recorded and were not found to be significantly different except where noted. Statistical tests were performed using Graphpad Prism and MATLAB.

Sholl analysis. VIP neurons were filled with a patch pipette containing 1% Alexa 647 Hydrazide and the internal solution was allowed to dialyze for at least 30 min before slices were fixed in 4% paraformaldehyde for 1 h at room temperature. Slices were then washed three times for 30 min in PBS before slices were mounted in Fluormount-G (Southern Biotech). Images were acquired using a Zeiss Axio Imager microscope with a 20× objective with the use of an Apotome (Zeiss). Neurons were reconstructed using NeuronJ (ImageJ), and Sholl analysis was performed using a custom script in MATLAB.

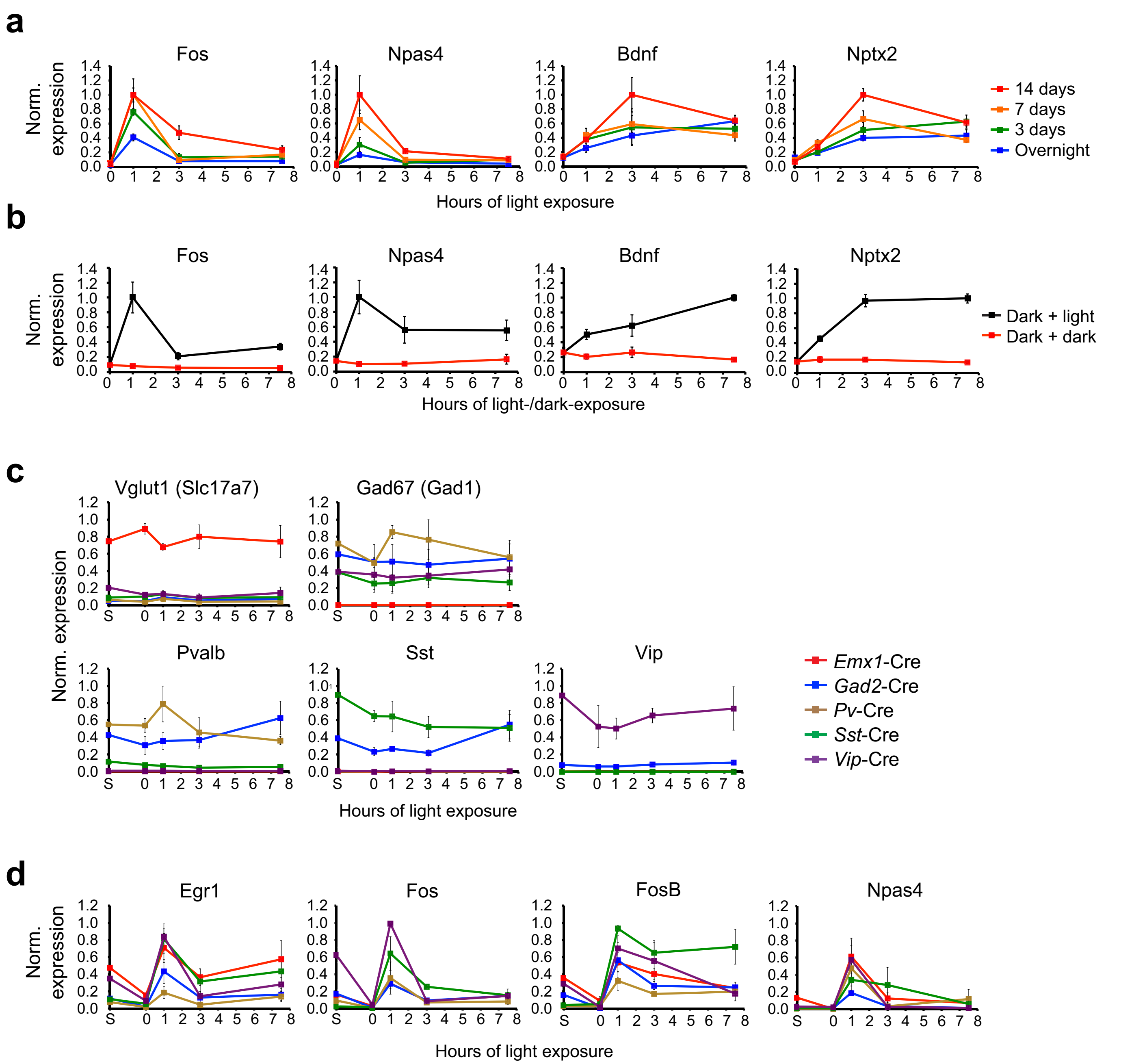
Monocular deprivation (MD) procedure. Eyelids were trimmed and sutured under isoflurane anaesthesia (1–2% in O_2) as previously described[31]. The integrity of the suture was checked daily and mice were used only if the eyelids remained closed throughout the duration of the deprivation period. One eye was closed for 4 days starting between P26 to P28. The eyelids were reopened immediately before recording, and the pupil was checked for clarity.

Mouse visual-evoked potential (VEP). VEPs were recorded from anaesthetized mice (50 mg kg^{-1} Nembutal and 0.12 mg chlorprothixene) using standard techniques described previously[32]. The contra- and the ipsilateral eye of the mouse were presented with horizontal black and white sinusoidal bars that alternated contrast (100%) at 2 Hz. A tungsten electrode was inserted into the binocular visual cortex at 2.8 mm from the midline where the visual receptive field was approximately 20° from the vertical meridian. VEPs were recorded by filtering the signal from 0.1–100 Hz and amplifying 10,000 times. VEPs were measured at the cortical depth where the largest amplitude signal was obtained in response to a 0.05 c.p.d. stimulus (400–600 μm); 3–4 repetitions of 20 trials each were averaged in synchrony with the abrupt contrast reversal. The signal was baseline corrected to the mean voltage of the first 50 ms post-stimulus-onset. VEP amplitude was calculated by finding the minimum voltage (negative peak) within a 50–150 ms post-stimulus-onset time window. Acuity was calculated only from the deprived eye. For each different spatial frequency, 3–4 repetitions of 20 trials each were averaged in synchrony with the abrupt contrast reversal. VEP amplitude was plotted against the log of the different spatial frequency, and the threshold of visual acuity was determined by linear extrapolation to 0 μV.

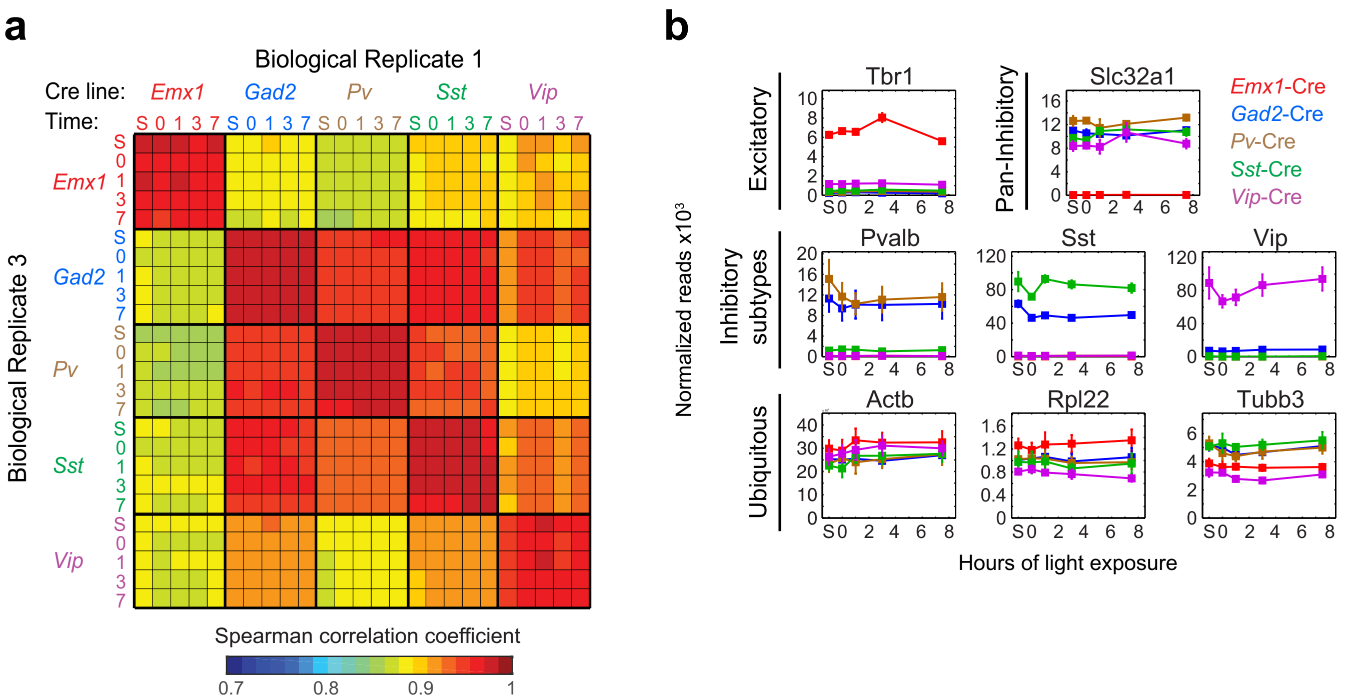
Animal husbandry and colony management. *Igf1* conditional knockout mice[15], Ai9 tdTomato reporter mice[33], *Emx1-cre*[34], *Pv-cre*[35], *Gad2-cre*, *Sst-cre*, *Vip-cre*[36] and RiboTag mice[8] are available from The Jackson Laboratory.

For routine experimentation, animals were genotyped using a PCR-based strategy; PCR primer sequences are available at the The Jackson Laboratory's website. For RiboTag experiments, mice homozygous for the RiboTag allele were crossed to mice homozygous for the *cre* allele and all experiments were performed with mice double heterozygous for both the RiboTag and the *cre* alleles. For *Igf1* cKO experiments, mice heterozygous for the *Igf1* conditional allele ($Igf1^{fl/WT}$) and homozygous for the *Vip-cre* allele were crossed to mice heterozygous for the *Igf1* conditional allele and homozygous for the tdTomato reporter allele. Resulting littermates all had one copy of the *Vip-cre* transgene and the tdTomato Cre reporter and yielded $Igf1^{WT/WT}$ and $Igf1^{fl/fl}$ littermates for experimentation. For injections of AAV constructs in the visual cortices of Cre mice (*Vip-*, *Pv-*, *Sst-*, or *Emx1-cre*), mice homozygous for the *cre* allele were crossed to wild-type C57Bl6 mice and offspring heterozygous for the *cre* allele were used for experiments. The use of animals was approved by the Animal Care and Use Committee of Harvard Medical School and/or the University of California Berkeley.

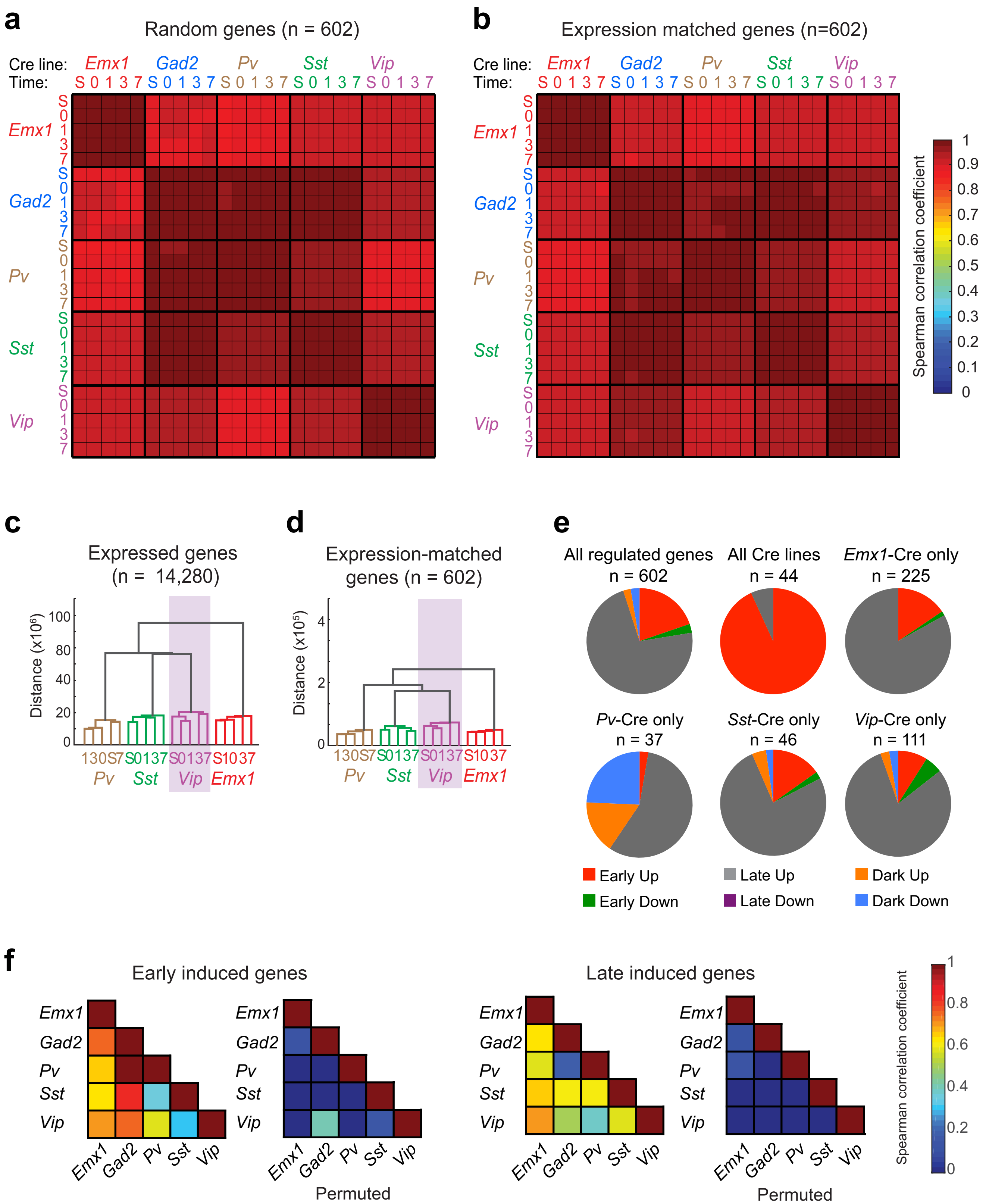
24. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* **25,** 402–408 (2001).
25. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (2012).
26. Rubinson, D. A. *et al.* A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. *Nature Genet.* **33,** 401–406 (2003).
27. Schopman, N. C. T., Liu, Y. P., Konstantinova, P., ter Brake, O. & Berkhout, B. Optimization of shRNA inhibitors by variation of the terminal loop sequence. *Antiviral Res.* **86,** 204–211 (2010).
28. Spiegel, I. *et al.* A central role for Necl4 (SynCAM4) in Schwann cell-axon interaction and myelination. *Nature Neurosci.* **10,** 861–869 (2007).
29. Schaeren-Wiemers, N. & Gerfin-Moser, A. A single protocol to detect transcripts of various types and expression levels in neural tissue and cultured cells: in situ hybridization using digoxigenin-labelled cRNA probes. *Histochemistry* **100,** 431–440 (1993).
30. Tiscornia, G., Singer, O. & Verma, I. M. Production and purification of lentiviral vectors. *Nature Protocols* **1,** 241–245 (2006).
31. Gordon, J. A. & Stryker, M. P. Experience-dependent plasticity of binocular responses in the primary visual cortex of the mouse. *J. Neurosci.* **16,** 3274–3286 (1996).
32. Durand, S. *et al.* NMDA receptor regulation prevents regression of visual cortical function in the absence of Mecp2. *Neuron* **76,** 1078–1090 (2012).
33. Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nature Neurosci.* **13,** 133–140 (2010).
34. Gorski, J. A. *et al.* Cortical excitatory neurons and glia, but not GABAergic neurons, are produced in the Emx1-expressing lineage. *J. Neurosci.* **22,** 6309–6314 (2002).
35. Hippenmeyer, S. *et al.* A developmental switch in the response of DRG neurons to ETS transcription factor signaling. *PLoS Biol.* **3,** e159 (2005).
36. Taniguchi, H. *et al.* A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. *Neuron* **71,** 995–1013 (2011).



Extended Data Figure 1 | Validation of the sensory stimulation protocol and the RiboTag-based cell-type-specific purification of mRNA. **a**, Quantitative real-time PCR (qPCR) for known experience-regulated genes on RNA isolated from the visual cortex of mice that were dark-housed for varying durations (overnight, 3 days, 7 days or 14 days) and then either euthanized in the dark or exposed to light for 1, 3 or 7.5 h, and then euthanized. Data are normalized to the maximal value in each data set and represent the mean and standard error of four biological replicates. **b**, qPCR for known experience-regulated genes on RNA isolated from the visual cortex of mice that were dark-housed for 14 days and then either exposed to light for 1, 3 or 7.5 h (dark + light, black) or kept in the dark during these hours (dark + dark, red). All mice of a given time point were dissected in very close temporal proximity. Data are normalized to the maximal value in each data set and represent the mean and standard error of four biological replicates. **c**, qPCR for known cell-type-specific marker genes on RNA isolated from RiboTag mice expressing Cre in distinct neuronal subtypes. Data are normalized to the maximal value in each data set and represent the mean and standard error of three biological replicates. **d**, qPCR for known early-induced transcription factors on RNA isolated from RiboTag mice expressing Cre in distinct neuronal subtypes. Data are normalized to the maximal value in each data set and represent the mean and standard error of three biological replicates.

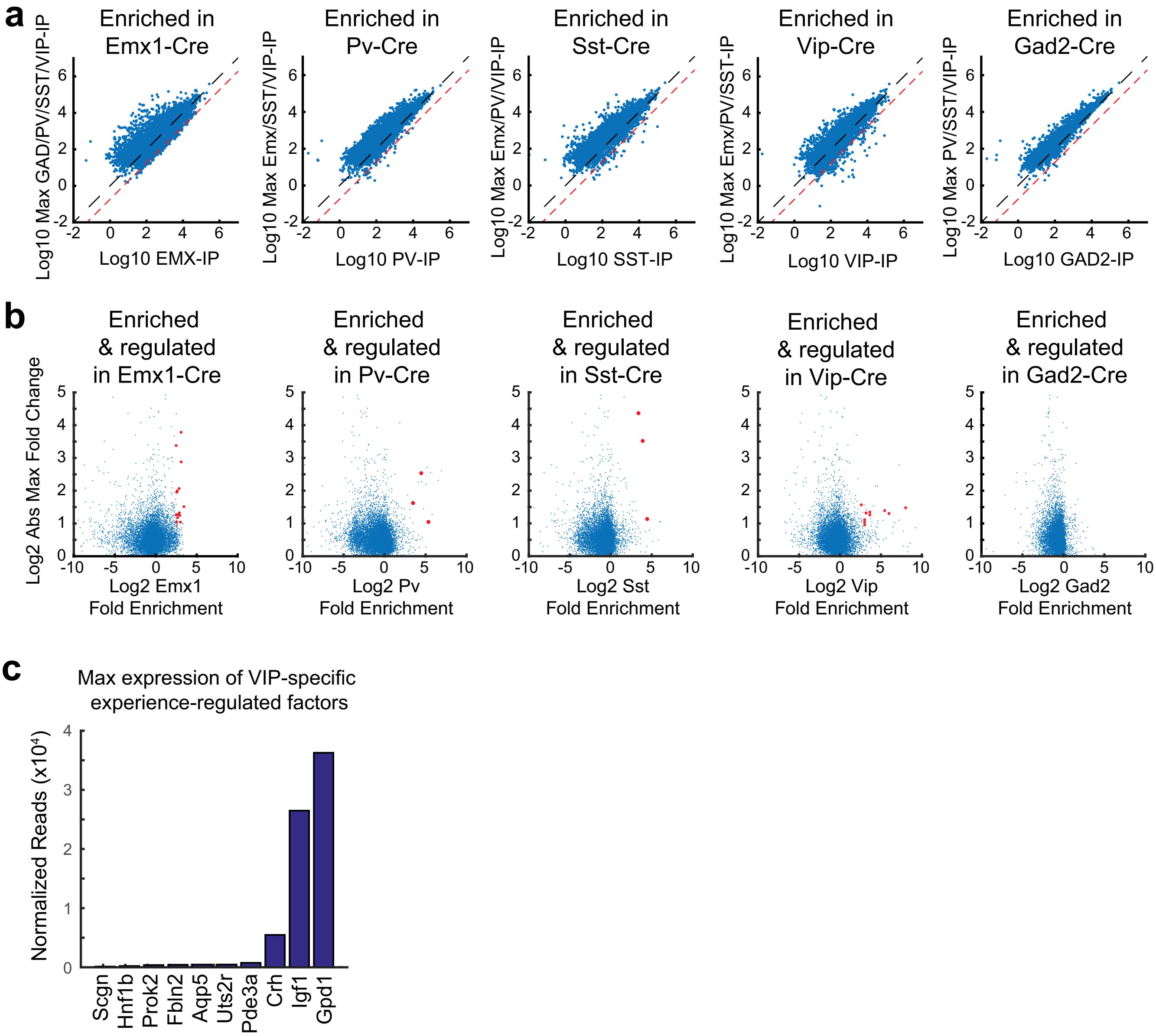


Extended Data Figure 2 | Validation of the RiboTag-seq approach. **a**, Matrix of Spearman correlation coefficients between biological replicates across all samples (scale of correlation coefficients extends from 0.7 to 1, see colour bar) (S, standard housing; 0, dark-housed only; 1/3/7.5, 1/3/7.5 h of light exposure after dark housing, respectively). **b**, Line plots of RNA-seq data showing the expression values (normalized reads across all exons of a gene) for cell-type-specific marker genes and ubiquitously expressed house-keeping genes in different Cre lines (*Emx1-cre*, red; *Gad2-cre*, blue; *Pv-cre*, brown; *Sst-cre*, green; *Vip-cre*, purple) across all time points of the experiment. Data represent the mean and standard error of three biological replicates.

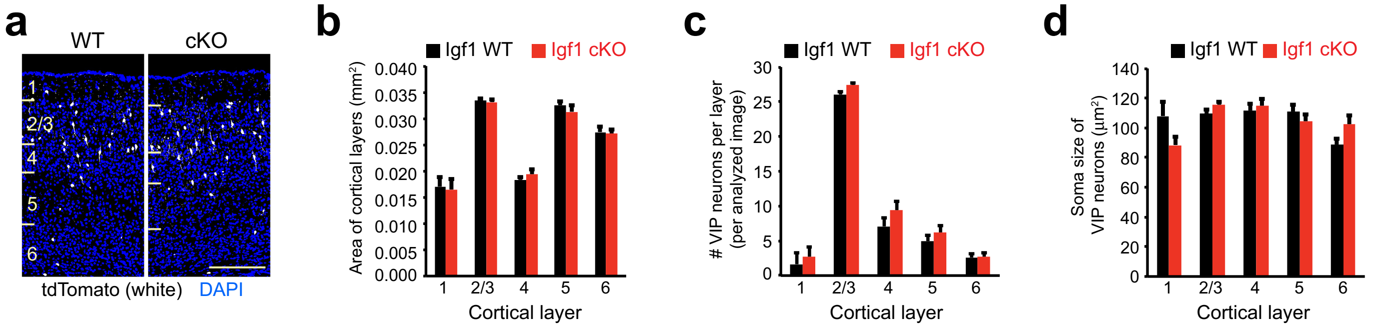


Extended Data Figure 3 | See next page for caption.

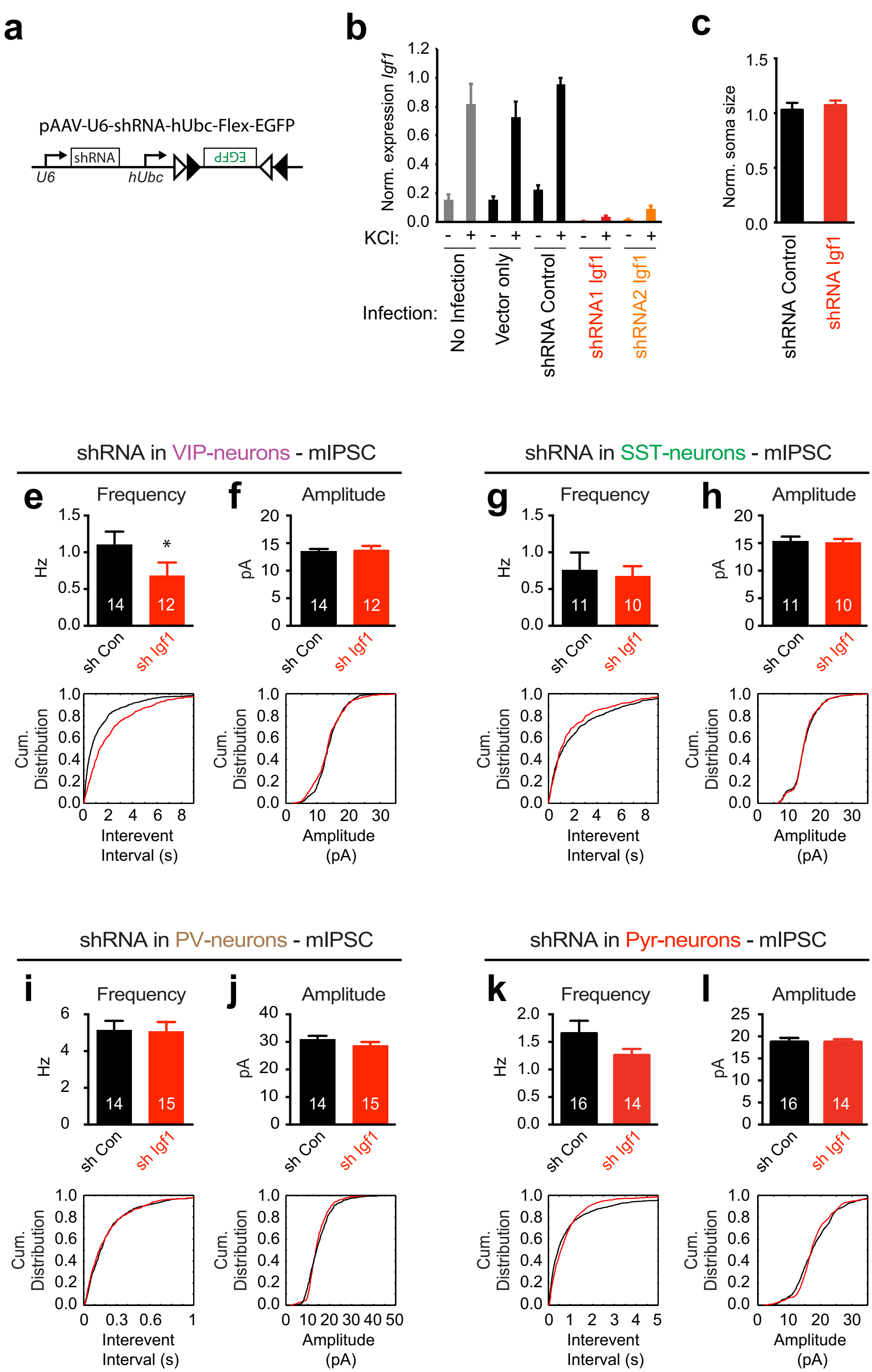
Extended Data Figure 3 | Characterization of the experience-induced gene programs in subtypes of cortical neurons. a, Average matrix of Spearman correlation coefficients computed from the expression values of 1000 random sets of 602 genes (including experience-regulated genes, with replacement). **b**, Matrix of the Spearman correlation coefficients computed from the expression levels of control transcripts that match the expression distribution of experience-regulated genes ($n = 602$). **c**, Cladogram resulting from hierarchical clustering of all samples (except samples from *Gad2-cre*). Cladograms were computed using the mean expression values (that is, normalized reads across all exons of a gene) for all expressed transcripts ($n = 14,280$). **d**, Cladogram resulting from hierarchical clustering of the mean expression values of a set of control transcripts that match the expression distribution of experience-regulated genes ($n = 602$). **e**, Pie charts showing the subdivision of experience-regulated genes on the basis of kinetics in each set of Cre lines (red, rapidly induced; grey, induced with delayed kinetics; orange, induced only after two weeks of dark housing; green, rapidly suppressed; magenta, suppressed with delayed kinetics; blue, suppressed only after two weeks of dark housing). **f**, Left, matrix of Spearman correlation coefficients between Cre lines computed using the mean expression values (normalized reads across all exons of a gene) of early-induced genes one hour after light exposure. Right, matrix of Spearman correlation coefficients between Cre lines computed using the mean expression values of late-induced genes 7.5 h after light exposure. For each matrix, the correlations upon permuting the expression values are also shown (colour scale at right, scale begins at zero.)



Extended Data Figure 4 | Characterization of cell-type-specific and experience-induced genes in subtypes of cortical neurons. a, Scatter plots showing the $\log_{10}$ expression values for each expressed gene in a given Cre line (*x* axis) plotted against the maximum $\log_{10}$ expression values for that gene found in all other Cre lines (*y* axis). Black line denotes unity, and the red line is the 5.5-fold enrichment threshold set to include *Vglut1* as a cell-type-specific gene in *Emx1-cre*. Data represent the mean values of three biological replicates. **b**, Scatter plots of all expressed genes, for each Cre line plotting the mean $\log_2$ fold enrichment in that Cre line (*x* axis) against the mean $\log_2$ of the absolute value of the maximum fold change observed in that Cre line. Data represent the mean values of three biological replicates. Genes that pass both enrichment and induction thresholds in 3 of 3 biological replicates are shown in red. **c**, Bar graph showing the maximum expression value (in normalized reads) for VIP-neuron-specific experience-regulated genes.

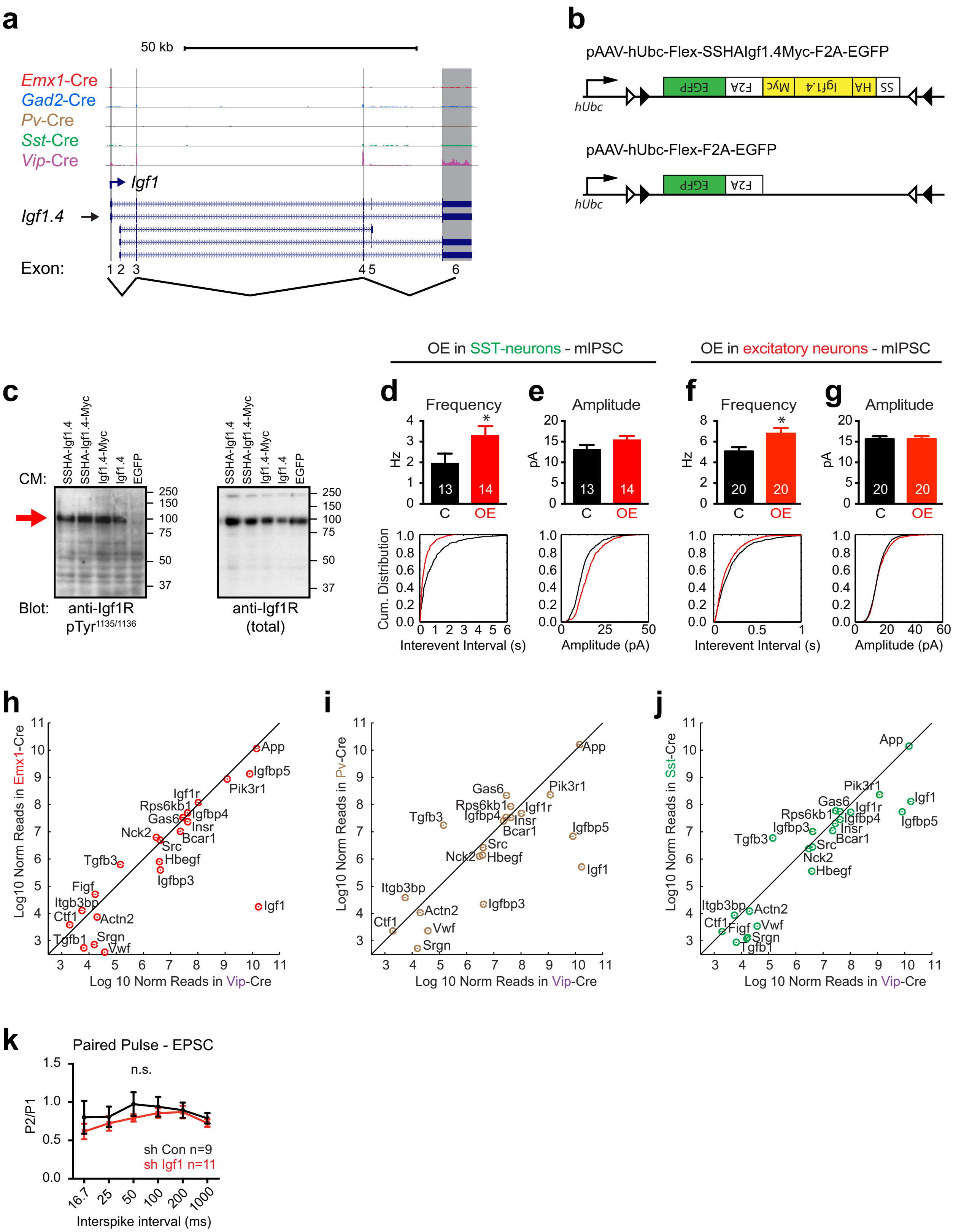


Extended Data Figure 5 | Conditional knockout of *Igf1* in VIP neurons does not affect cortical morphology or gross morphology of VIP neurons. a, Example image of cortices from *Igf1* wild-type (WT) (*Vip-cre*/+, *LSL-tdTomato*/+, $Igf1^{WT/WT}$) or conditional-knockout (cKO) (*Vip-cre*/+, *LSL-tdTomato*/+, $Igf1^{fl/fl}$) mice. VIP neurons are labelled in white, with DAPI shown in blue (cortical layers are indicated on the left, scale bar, 200 µm). **b**–**d**, Bar graphs showing the area of each cortical layer (**b**), number of VIP neurons per image per layer (**c**), or soma size of VIP neurons (**d**) in *Igf1* wild-type (black) or conditional-knockout (red) mice. Data represent the mean and standard error of three biological replicates.



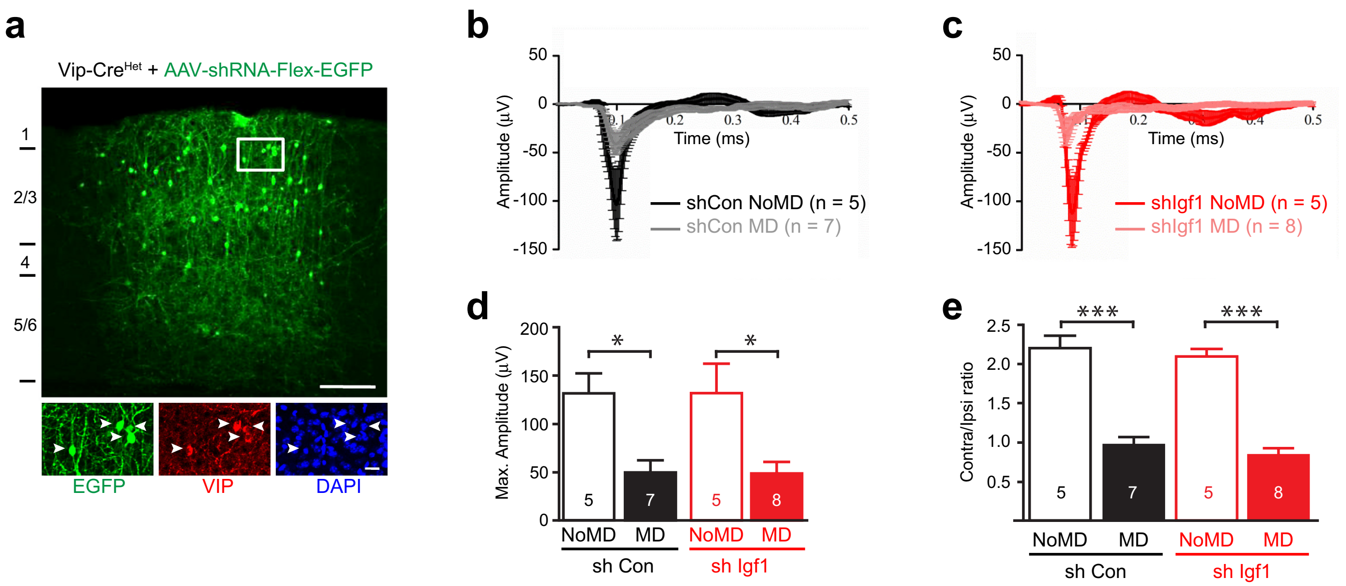
Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | *Igf1* knockdown in VIP neurons affects inhibitory inputs onto VIP neurons but not onto neighbouring neurons. a, AAV shRNA constructs. shRNA cassettes against *Igf1* or a control gene (*Luc*) were cloned downstream of the U6 promoter into an AAV vector that drives Cre-dependent expression of eGFP. **b**, qPCR validation of the efficacy of *Igf1* shRNA constructs. Cultured cortical neurons were infected with lentiviral constructs either expressing no shRNA (vector only), a control shRNA (against *Luc*) or shRNAs against *Igf1*. Four days post-infection the cultures were quieted overnight with TTX and AP-5 and then harvested either before or after being depolarized with 55 mM KCl for 6 h; RNA was then isolated and qPCR was performed. Data are normalized to the maximal value in each replicate and represent the mean and standard error of three biological replicates. **c**, Bar graph showing normalized soma size of P21 visual cortex VIP neurons infected with control shRNA or shRNA targeting *Igf1* (shRNA control, $n = 103$; shRNA *Igf1*, $n = 174$; $P = 0.41$, Mann–Whitney *U*-test). **d**, Bar graphs showing the levels of IGF1 in the serum of P20 mice that were injected intracortically with AAVs driving the expression of control shRNA (black) or *Igf1* shRNA. Data represent the mean and s.e.m. of the serum IGF1 levels of four mice per group. **e**–**l**, Bar graphs and cumulative distribution plots showing mIPSC amplitudes and frequency/inter-event interval upon early widespread knockdown of *Igf1* in VIP (**e**, **f**), SST (**g**, **h**), PV (**i**, **j**) and excitatory (**k**, **l**) neurons after injection of AAVs into P3 cortices of the respective Cre mice. VIP neurons (identified as eGFP-positive cells in *Vip-cre* mice): control and *Igf1* shRNA, amplitude $P = 0.96$, frequency $P = 0.04$. SST neurons (identified as eGFP-positive cells in *Sst-cre* mice): control and *Igf1* shRNA, amplitude $P = 0.89$, frequency $P = 0.55$. PV neurons (identified as eGFP-positive cells in *Pv-cre* mice): control and *Igf1* shRNA, amplitude $P = 0.084$, frequency $P = 0.93$. Pyramidal neurons (identified by morphology): control and *Igf1* shRNA, amplitude $P = 0.84$, frequency $P = 0.15$). All *P* values are derived from Mann–Whitney *U*-tests; numbers inside bars indicate the number of cells recorded.



Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Effects of IGF1 overexpression in excitatory and SST-positive neurons. **a**, RiboTag-seq identifies *Igf1.4* as the major *Igf1* isoform expressed in VIP neurons. Representative tracks of histograms of the RNA-seq reads in each Cre line across the *Igf1* genomic locus. Data are from the 7.5 h light exposure RiboTag-seq data sets. **b**, AAV constructs for the Cre-dependent expression of HA-/Myc-tagged IGF1 (*Igf1.4*) and eGFP (that is, IGF1-OE, top) or of eGFP alone (that is, control, bottom). F2A, Furin cleavage site followed by the 2A peptide; black and white triangles represent a Cre-dependent Flex-switch. **c**, Western blot analysis of IGF1-receptor activation status in lysates of serum starved HEK293T cells that were stimulated with conditioned media (CM) containing epitope-tagged isoforms of IGF1. CM was produced by transfecting HEK293T cells with the respective construct and collecting the culture media. IGF1-receptor is detected with antibodies against either activated IGF1 receptor (anti-IGF1R pTyr$^{1136/1138}$) or total IGF1 receptor (anti-IGF1R total). Molecular weight markers are on the right and the arrow indicates the band of the IGF1 receptor. **d**, **e**, Bar graphs and cumulative distribution plots showing mIPSC frequency/inter-event interval (**d**) and amplitude (**e**) of mIPSCs recorded from eGFP-positive neurons in P20 *Sst-cre* mice that were intracortically injected with AAVs driving the expression of control (C) or IGF1-OE (OE) constructs (amplitude, $P = 0.16$; frequency, $P = 0.01$; Mann–Whitney *U*-test; numbers inside bars indicate the number of cells recorded). **f**, **g**, Bar graphs and cumulative distribution plots showing mIPSC frequency/inter-event interval (**f**) and amplitude (**g**) of mIPSCs recorded from eGFP-positive neurons in P20 *Emx1-cre* mice that were intracortically injected with AAVs driving the expression of control (black, $n = 20$) or IGF1-OE AAVs (red, $n = 20$). Amplitude, $P = 0.99$; frequency, $P = 0.01$, Mann–Whitney *U*-test. **h**–**j**, Scatter plots of IGF1-interacting proteins showing the $\log_{10}$ normalized mean expression values in *Vip-cre* neurons versus each of the other Cre lines (*Emx1* (**h**), *Pv* (**i**), *Sst* (**j**)). **k**, Quantification of EPSC paired-pulse recordings from VIP neurons infected with control shRNA- (black $n = 9$) or *Igf1* shRNA- (red $n = 11$) expressing AAVs. The ratio of the second EPSC amplitude divided by the first EPSC amplitude is plotted against inter-stimulus interval ($P = 0.1$, two-way ANOVA).



Extended Data Figure 8 | VIP neuron-derived IGF1 does not disrupt ocular dominance plasticity. a, Widespread infection of VIP neurons by AAV-shRNA-hUbc-Flex-eGFP. High-titre injection of AAVs into the visual cortex of P18–20 *Vip-cre/+* mice leads to infection of the majority VIP neurons (green, eGFP; red, anti-VIP; blue, DAPI; arrowheads, infected VIP neurons; scale bars, 150 µm, 20 µm in the inlet). **b**, **c**, Average of VEP traces recorded in the visual cortices of mice that were injected with AAVs expressing control shRNA (black/grey) or shRNA against *Igf1* (red/pink) shRNA and that were (grey, pink) or were not (black, red) subjected to monocular deprivation in the eye contralateral to the recording site (MD versus NoMD, respectively). **d**, Monocular deprivation induces a significant reduction in the VEP amplitude in response to low spatial frequency stimulation in mice that had AAVs expressing control shRNA and *Igf1* shRNA injected into their visual cortices (control shRNA NoMD, $n = 5$ mice; control shRNA MD, $n = 7$ mice; *Igf1* shRNA NoMD, $n = 5$; *Igf1* shRNA MD, $n = 8$. $*P < 0.05$, Mann–Whitney *U*-test). **e**, Mice that had AAVs expressing control shRNA (black) and *Igf1* shRNA (red) injected into their visual cortices display normal ocular dominance plasticity as monocular deprivation (MD) induces a shift to the ipsilateral eye in both groups (control shRNA NoMD, $n = 5$ mice; control shRNA MD, $n = 7$; *Igf1* shRNA NoMD, $n = 5$; *Igf1* shRNA MD, $n = 8$ mice. $**P < 0.0001$, one-way ANOVA with Tukey's post hoc test).

doi:10.1038/nature17000

Co-ordinated ocular development from human iPS cells and recovery of corneal function

Ryuhei Hayashi[1,2], Yuki Ishikawa[2], Yuzuru Sasamoto[2], Ryosuke Katori[2], Naoki Nomura[2], Tatsuya Ichikawa[2], Saori Araki[2], Takeshi Soma[2], Satoshi Kawasaki[2], Kiyotoshi Sekiguchi[3], Andrew J. Quantock[4], Motokazu Tsujikawa[2] & Kohji Nishida[2]

The eye is a complex organ with highly specialized constituent tissues derived from different primordial cell lineages. The retina, for example, develops from neuroectoderm via the optic vesicle, the corneal epithelium is descended from surface ectoderm, while the iris and collagen-rich stroma of the cornea have a neural crest origin. Recent work with pluripotent stem cells in culture has revealed a previously under-appreciated level of intrinsic cellular self-organization, with a focus on the retina and retinal cells[1–5]. Moreover, we and others have demonstrated the *in vitro* induction of a corneal epithelial cell phenotype from pluripotent stem cells[6–9]. These studies, however, have a single, tissue-specific focus and fail to reflect the complexity of whole eye development. Here we demonstrate the generation from human induced pluripotent stem cells of a self-formed ectodermal autonomous multi-zone (SEAM) of ocular cells. In some respects the concentric SEAM mimics whole-eye development because cell location within different zones is indicative of lineage, spanning the ocular surface ectoderm, lens, neuro-retina, and retinal pigment epithelium. It thus represents a promising resource for new and ongoing studies of ocular morphogenesis. The approach also has translational potential and to illustrate this we show that cells isolated from the ocular surface ectodermal zone of the SEAM can be sorted and expanded *ex vivo* to form a corneal epithelium that recovers function in an experimentally induced animal model of corneal blindness.

To generate a SEAM of ocular cells, human induced pluripotent stem (iPS) cells were cultivated in differentiation medium in which they spontaneously and progressively formed a primordium comprising four identifiable concentric zones (Fig. 1a, Extended Data Fig. 1a). Cell morphology in each zone was distinctive, creating a visible delineation between zones (Extended Data Fig. 1b). The innermost central area (zone 1) formed first and this was followed by the emergence of three more radially distant concentric cell populations; zones 2–4. (Fig. 1b, Supplementary Video). In our experiments $7.7 \pm 1.8\%$ of human iPS cells formed colonies and $67.9 \pm 4.9\%$ of these resulted in the generation of a SEAM ($n = 5$ technical replicates). Immunolabelling for the neural cell marker class III β-tubulin (TUBB3) was positive in zones 1 and 2, but not, more peripherally, in zones 3 or 4 (Fig. 1c). Cells in zones 1–3 expressed the ocular cell marker PAX6, while those in zones 3 and 4 were positive for the epithelial/surface ectodermal markers p63 and E-cadherin (Fig. 1d, e, f). Thus, in a number of respects SEAM formation in two-dimensions mirrors whole eye development from the front of the ocular surface posteriorly to the retina (Fig. 1g).

Cells in zone 1 expressed neural cell-specific markers *TUBB3*, *SOX2* and *SOX6*, but no surface ectodermal markers (Fig. 2a). Accordingly, this central area was deemed to represent presumptive neuroectoderm. Zone 2 cells primarily expressed the optic vesicle marker *RAX* and the neural crest marker *SOX10* (Fig. 2a). In eye development various cell types assume a tissue-specific arrangement at the juxtaposition of the embryonic anterior and posterior eye segments (Extended Data Fig. 1c), and perhaps this is reflected here. We also note that zone 2 frequently contained retinal pigment epithelium (RPE)-like cells, and that $SOX10^{+}$/$p75^{+}$ neural crest cells were found to have emerged in satellite spheres in the presumptive zone 2 after two weeks in culture (Fig. 2b, Extended Data Fig. 1d). Differentiation of SEAMs by a protocol which encourages retinal differentiation further demonstrated that $CHX10^{+}$ neuro-retinal cells and $MITF^{+}$ RPE cells were present in zone 2 towards its inner and outer margins, respectively (Fig. 2c, d). Collectively, the findings imply that zone 2 of the ocular SEAM is a developmental analogue of neural eye tissues comprising the neuro-retina, neural crest and RPE.

At the margin of zones 2 and 3 α-crystallin^{+} lens cell clusters emerged after four weeks in culture and spread further through the SEAM by week six (Fig. 2e, Extended Data Fig. 1e). Cells in zone 3 did not display any neural features and this region was unique with its PAX6/p63-double-positive phenotype, representative of ocular surface ectoderm. Zone 3 cells also specifically expressed ocular surface ectoderm markers such as *PAX6*, *deltaN (DN)-p63*, *K18*, and *E-cadherin*, but no neural cell markers (Fig. 2a). Thus, cells in zone 3 are considered to be anlages of the ocular surface epithelium. Cells in zone 4 expressed epithelial genes *DN-p63* and *E-cadherin*, and did not express *PAX6*. This points to their identity as general surface ectodermal cells, which will probably differentiate into epidermal keratinocytes. The interactions of the different cell lineages complicit in whole eye formation *in situ* are thus mimicked in the research described here as illustrated schematically in Figs 1g and 2f.

As mentioned, cells with retina-like characteristics have been generated from human iPS cells[1–5], but functional ocular surface tissue has not. Thus, we sought to form a transplantable corneal epithelium; (i) as a conceptual example of the translational potential of the SEAM and (ii) to demonstrate that functional anterior eye tissue can indeed be fashioned from human iPS cells (Fig. 3a). At two and four weeks in culture cells in zone 3 co-expressed PAX6 and p63, first partially then fully. After isolation by the manual pipetting of cells in other zones (Extended Data Fig. 2a, b), zone 3 cells were also found to express K14 and cornea-specific keratin K12[10] by 8–12 weeks (Fig. 3b). This was confirmed by qRT–PCR analysis (Extended Data Fig. 2c). Vimentin positive (VIM^{+}) stroma-like cells were also present in the $p63^{+}$ zone 3 at eight weeks (Extended Data Fig. 2d). Cellular differentiation in zone 3 of the SEAM thus follows that seen in the embryonic (mouse) eye (Extended Data Fig. 2e). BMP signalling influences the development of the surface ectoderm[11,12], and we found that zone 3 of the SEAM was abolished following its treatment with the BMP4 inhibitors Noggin and LDN-193189 (Extended Data Fig. 3). The TGFβ inhibitor, SB-431542, also disturbed the typical multi-zone formation (Extended Data Fig. 3), suggesting that the suppression of ocular

[1]Department of Stem Cells and Applied Medicine, Osaka University Graduate School of Medicine, Suita, Osaka 565-0871, Japan. [2]Department of Ophthalmology, Osaka University Graduate School of Medicine, Suita, Osaka 565-0871, Japan. [3]Laboratory of Extracellular Matrix Biochemistry, Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan. [4]Structural Biophysics Group, School of Optometry and Vision Sciences, College of Biomedical and Life Sciences, Cardiff University, Cardiff CF24 4HQ, UK.

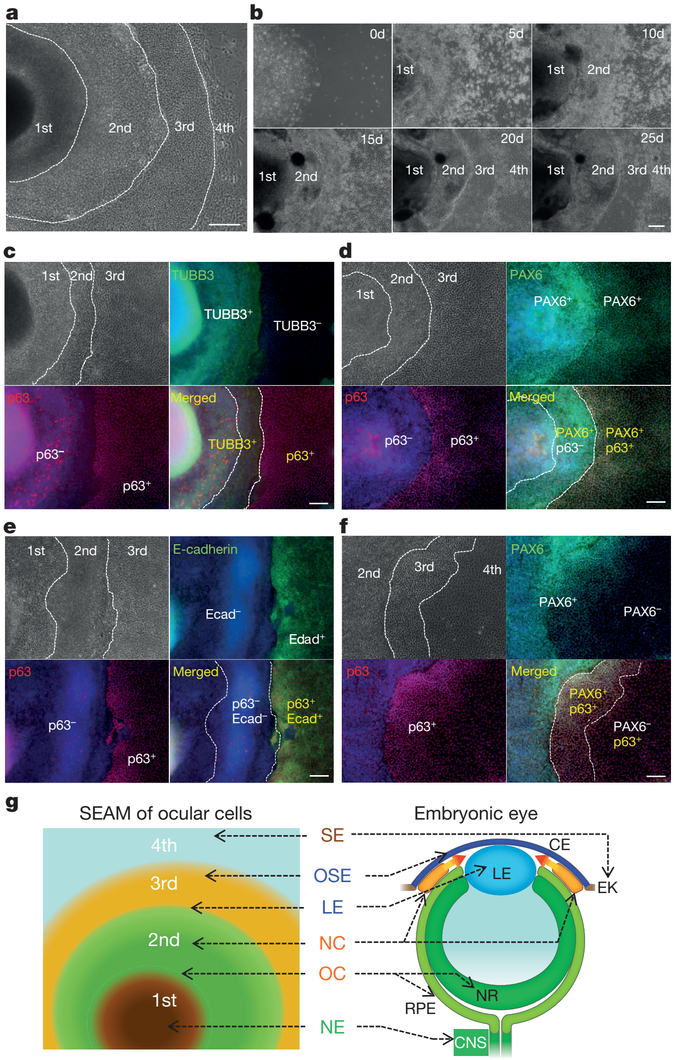


Figure 1 | Differentiation of multiple ocular cells in the SEAM. **a**, A typical SEAM of differentiated human iPS cells after 40 days of culture. Representative of 19 independent experiments. Scale bar, 200 μm. **b**, Time-lapse microscopy of the differentiating human iPS cells during the first 25 days (d) of culture. Representative of six independent experiments. Scale bar, 200 μm. See also Supplementary Video 1. **c**–**f**, Immunostaining of TUBB3 (green) and p63 (red) (**c**, zones 1–3), PAX6 (green) and p63 (red) (**d**, zones 1–3), E-cadherin (Ecad, green) and p63 (red) (**e**, zones 1–3), and PAX6 (green) and p63 (red) (**f**, zones 2–4) in the SEAM after six to seven weeks of culture. Images are representative of three or four independent experiments. Nuclei, blue. Scale bar, 100 μm. **g**, The SEAM of human iPS cells induced different kinds of cells of ectodermal lineage, mimicking anterior and posterior eye development *in vivo*. CNS, central nervous system; NE, neuroectoderm; OC, optic cup; NR, neuroretina; NC, neural crest; LE; lens; OSE, ocular surface ectoderm; SE, surface ectoderm; CE, corneal epithelium; EK, epidermal keratinocyte.

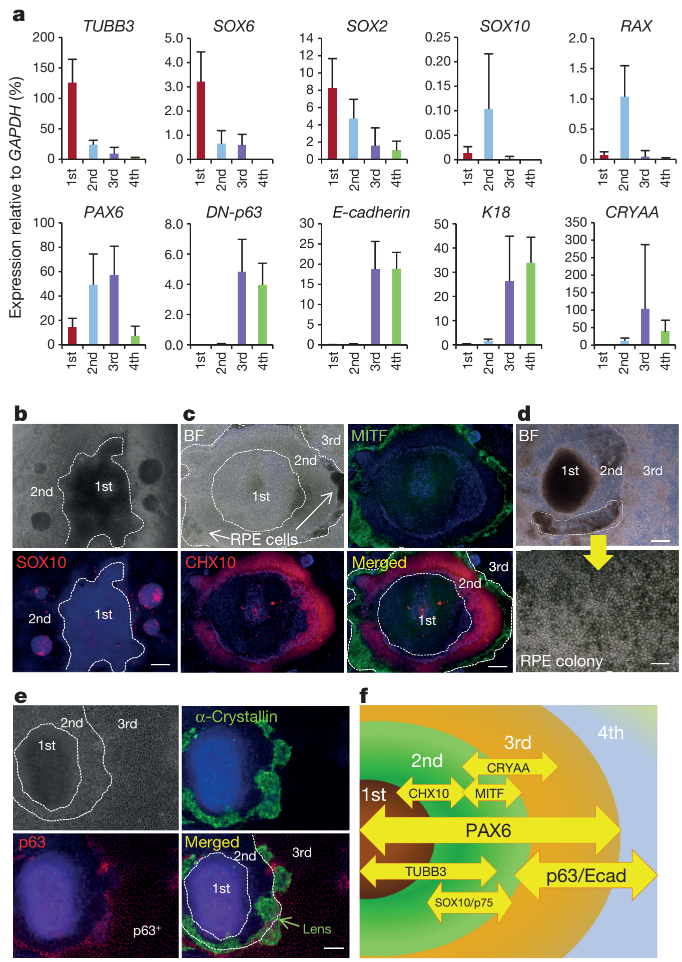


Figure 2 | Characterization of cell zones in the SEAM. **a**, Gene expression analysis for ectoderm-related markers in each zone of the SEAM (zones 1 and 2, $n = 12$ colonies; zones 3 and 4, $n = 7$ colonies obtained from two independent experiments, respectively). Error bars are s.d. **b**, Phase-contrast image and SOX10 (red) expression around SEAM zones 1 and 2 after two weeks of culture (representative of three independent experiments). Nuclei, blue. Scale bar, 200 μm. **c**, Bright-field (BF) appearance of neuroretinal cells (CHX10$^+$, red) and RPE cells (MITF$^+$, green) in zone 2 after seven weeks of culture using the protocol for retinal differentiation (representative of three independent experiments). Nuclei, blue. Scale bar, 200 μm. **d**, Cultivation of the isolated RPE cells (representative of three independent experiments). Scale bars, 200 μm (upper panel) and 50 μm (lower panel). **e**, Immunostaining of lens differentiation marker, α-crystallin (green) and p63 (red), after four weeks of culture (representative of three independent experiments). Nuclei, blue. Scale bar, 100 μm. **f**, Schematic of the expression pattern of ocular-cell-related genes in the SEAM.

surface ectodermal commitment is caused by the inhibition of early developmental events induced by endogenous BMP/TGFβ.

As alluded to above, cells in SEAM zone 3 most closely resemble those of the presumptive ocular surface, and these were investigated for their translational, regenerative potential. After removal of zones 1 and 2 from the SEAM by manual pipetting, the cells that remained (that is, those in zone 3 and some in zone 4) were subjected to FACS to isolate a specific ocular surface lineage—corneal or conjunctival—and the results of this are shown in Extended Data Fig. 4a. SSEA-4 is a common marker of pluripotent stem cells and is specifically expressed in corneal epithelial cells *in vivo*[13], including stem/progenitor cells (Extended Data Fig. 4b). Here, the use of SSEA-4 and the basal epithelial marker ITGB4 revealed that 14.1% of cells were SSEA-4$^+$/ITGB4$^+$ (P3 cells), whereas 16.6% were SSEA-4$^-$/ITGB4$^+$ (P2 cells) (Fig. 3c). Thus, the population contains both corneal and non-corneal epithelial cells, with the indication that most colony-forming cells are derived from P2 and P3 fractions (Fig. 3d). Unlike P2, however, a significant proportion of the P3 colonies expressed PAX6 (Extended Data Fig. 4c). P3 cells also had higher expression levels of corneal epithelial-specific markers *PAX6*, *K12*, *CLU*, and *ALDH3A1*[14], but lower levels of the epidermal marker *K10* and the mucosal epithelial marker *K13* (Extended Data Fig. 4d).

P3 cell-derived colonies mostly consisted of PAX6$^+$/K12$^+$ corneal epithelial colonies along with a lower number of PAX6$^+$/K12^{low} limbal epithelial colonies, which are assumed to represent corneal epithelial stem/progenitor cells (Fig. 3e, Extended Data Fig. 5a). The limbus, at the edge of the cornea, is where corneal epithelial stem/progenitor cells are widely believed to reside. Stratified epithelia derived from

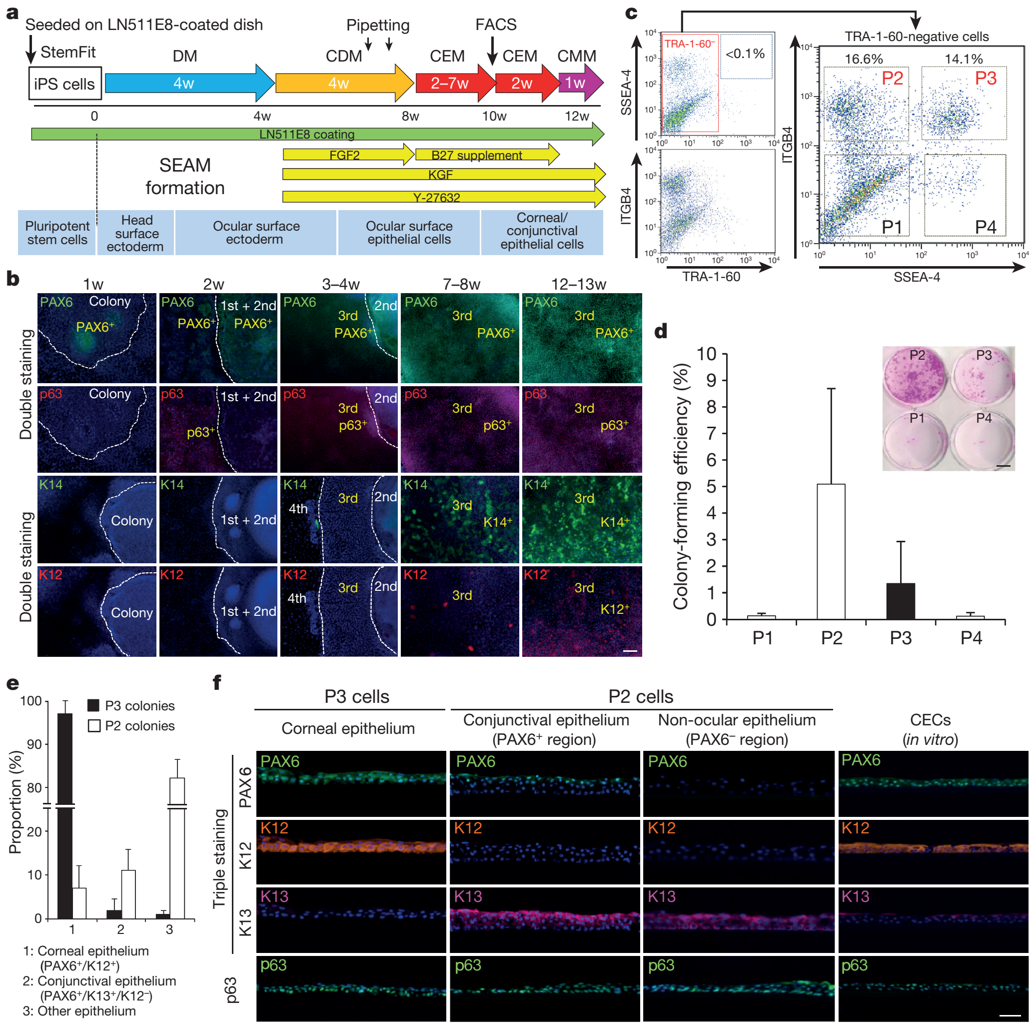


Figure 3 | Induction and isolation of ocular surface epithelial cells from the SEAM. a, Schematic representation of the strategy used for the generation of the ocular surface ectoderm and epithelium. CDM, corneal differentiation medium; CEM, corneal epithelium maintenance medium; CMM, corneal epithelium maturation medium; DM, differentiation medium; w, week(s). **b**, Immunostaining for ocular surface and corneal epithelial development-related markers, PAX6 (green), p63 (red), K14 (green), and K12 (red) during ocular surface epithelial differentiation culture (1–12 weeks, representative of three independent experiments). Nuclei, blue. Scale bar, 100 μm. **c**, A flow cytometric analysis of SSEA-4, ITGB4 and TRA-1-60 in the differentiated human iPS cells after 10–15 weeks of culture revealed no undifferentiated human iPS cells (SSEA-4^{+}/TRA-1-60^{+} cells) (left panels). TRA-1-60^{-} cells were analysed by SSEA-4 and ITGB4 (right panel). The four populations are defined as P1–P4. The data shown here are representative images of 23 independent cell-sorting experiments. **d**, Colony-forming assay for the sorted P1–P4 cell fractions. Fixed colonies were stained with rhodamine (right upper panel, representative of seven independent cell-sorting experiments) and the colony-forming efficiency were calculated (graph: seven independent cell-sorting experiments). Error bars are s.d. Scale bar, 5 mm. **e**, The ratios of corneal epithelial (PAX6^{+}/K12^{+} or PAX6^{+}/K12^{low}), conjunctival epithelial (PAX6^{+}/K13^{+}/K12^{-}) and other epithelial cell types in the P2 and P3 colonies from five independent cell sorting experiments. Error bars are s.d. **f**, Triple-colour immunostaining for PAX6 (green), K12 (orange), K13(magenta), and p63 (green) expression in stratified P2 and P3 cells and cultivated human corneal limbal epithelial cells (CECs) (representative of three independent experiments, respectively). Nuclei, blue. Scale bar, 50 μm.

P3 cells expressed K12 and PAX6, but not K13 (Fig. 3f) or *HOX* genes (Extended Data Fig. 5b). P2 cells, on the other hand, expressed *HOX* genes (Extended Data Fig. 5b), which are not expressed in cells of the ocular surface[15,16]. In the colonies of P2 cells, those expressing HOXB4 and PAX6 existed exclusively of each other (Extended Data Fig. 5c). Immunostaining of P2 cells further demonstrated that they consisted of PAX6^{+}/K13^{+} conjunctival epithelial cells[17] accompanied by a majority of PAX6^{-}/K13^{+} non-ocular epithelial cells (Fig. 3e, f, Extended Data Fig. 5a). Long-term differentiation revealed that PAS-stained, MUC5AC^{+}, K7^{+} conjunctival goblet-like cells appeared in presumptive P2 cell regions (Extended Data Fig. 5d). Collectively, our data indicates that cells in zone 3 of the SEAM have characteristics of corneal, limbal, and conjunctival epithelial cells (Supplementary Table 1), raising the possibility that functional ocular surface epithelia could be developed for surgical use.

To investigate this prospect we generated *ex vivo* expanded sheets of corneal epithelium from FACS-sorted cells acquired from zone 3 of the SEAM (Fig. 4a). The sheets expressed the corneal limbal stem-cell markers K15 and K19, along with the mucins MUC1, MUC4 and MUC16, tight junction protein ZO-1, the differentiation marker CX43, K3 and K12, all of which are characteristic for the cornea (Fig. 4b, Extended Data Fig. 6a). The non-corneal keratins K10 and K18 were not expressed (Fig. 4b). Approximately 99% of the cells in the expanded sheets were stratified K14^{+} epithelial cells, 95% were of corneal epithelial lineage (SSEA-4^{+}), and 70% were differentiated corneal epithelial cells (K12^{+}) (Fig. 4c). Cells typically had a smooth apical surface with microvilli-like structures (Extended Data Fig. 6b). Analyses based on significantly changed genes revealed that cells in the expanded corneal epithelial sheet differed from human oral mucosal keratinocytes, dermal fibroblasts and, indeed, human iPS

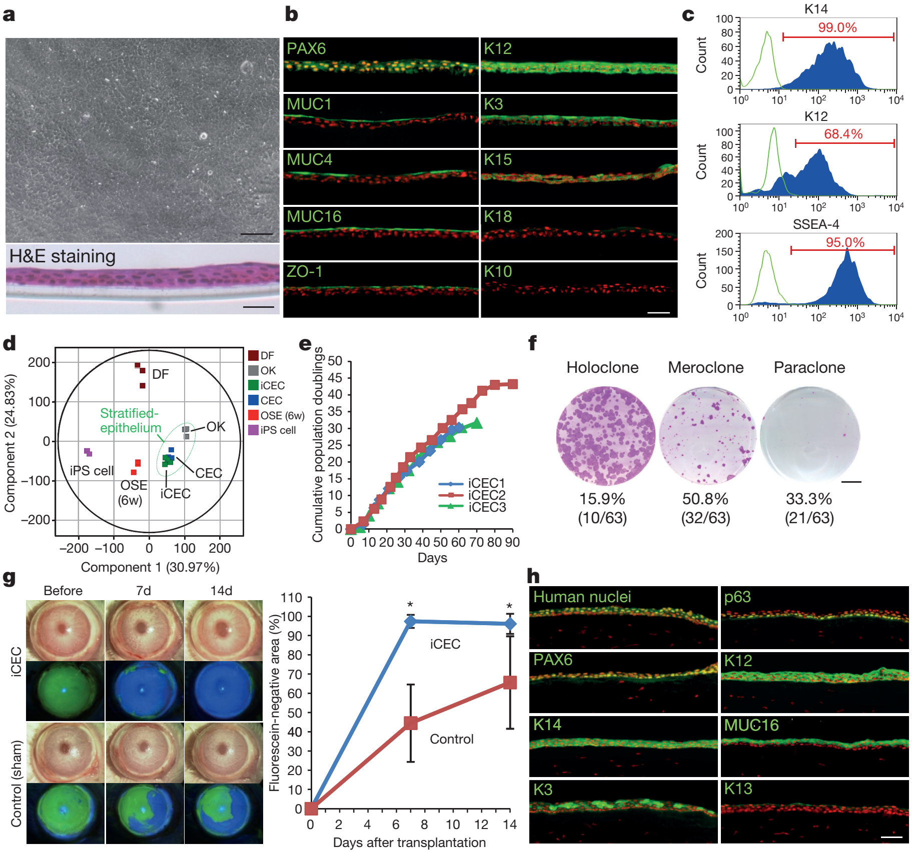


Figure 4 | Characterization and surgical use of human iPS cell-derived corneal epithelial cells (iCECs). a, Phase-contrast microscopy and haematoxylin and eosin (H&E) staining of the human iCECs on cell culture inserts (representative of four independent experiments, respectively). Scale bars, 100 μm (phase), 50 μm (H&E). **b**, Immunostaining for corneal epithelial functional proteins and epithelial markers (green) in the stratified SEAM-derived human iCEC sheets (representative of three independent experiments). Nuclei, red. Scale bar, 50 μm. **c**, Results of flow cytometric analyses for K14, K12 and SSEA-4 expression in the stratified human iCECs (representative of three independent experiments). **d**, Results of a principle component analysis based on the global gene expression (examined by microarrays) comparing human iPS cells ($n=3$ technical replicates), ocular surface ectoderm (OSE; that is, human iPS cell-derived cells after six weeks of differentiation, $n=3$ technical replicates), human oral keratinocytes (OKs, $n=3$ technical replicates), human iCECs ($n=4$ independent experiments), human dermal fibroblasts (DFs, $n=3$ technical replicates) and human corneal epithelial cells from the limbus (CECs, $n=3$ independent experiments). A total of 25,262 significantly changed genes (fold change >2.0, false discovery rate <0.05) were analysed. Compared to human DFs, human iPS cells, human OKs or the OSE, the gene expression in human iCECs was most similar to that of human CECs. **e**, Proliferation profiles of the human iCECs during serial passages ($n=3$ independent experiments, average population doubling, 35.1). **f**, Result of a holoclone analysis for human iCEC colonies. Representative images of holoclone-, meroclone- and paraclone-derived colonies and their frequencies are shown ($n=63$ single colonies from four independent cell-sorting experiments). Scale bar, 10 mm. **g**, Barrier function assay using fluorescein staining for transplanted and sham-operated control corneas on days 0, 7 and 14 post-surgery (left panels). A statistical analysis of the barrier function interpreted as the size of the fluorescein-negative area in the human iCEC-sheet-transplanted and control corneas (graph). *$P<0.05$; $n=7$ biological replicates, Steel's test (Bonferroni corrected). Error bars are s.d. **h**, Immunostaining for human nuclei and corneal epithelial markers (green) in the SEAM-derived human iCEC-sheet-transplanted corneas on postoperative day 14 ($n=6$ animal transplantation experiments). Nuclei, red. Scale bar, 50 μm.

cells, but were similar to cells harvested from the epithelium at the limbus of donated research human corneas (Fig. 4d, Extended Data Fig. 6c). Cells had the proliferative capability to be passaged 11 to 16 times (average population doubling, 35.1; Fig. 4e, Extended Data Fig. 6d) without any obvious karyotype aberration at each passage (Extended Data Fig. 6e). Holoclone analysis indicated that 15.9% of the corneal epithelial colony was composed of holoclones (Fig. 4f). Together, the data show that corneal epithelial cells derived from our SEAM contain functional stem/progenitor cells equipped with high proliferative potential, which have the potential to reconstruct anterior eye epithelial tissue. Furthermore, all five different human iPS cell clones used were able to successfully induce corneal epithelial stem/progenitor cells, though each had a differential propensity for engendering corneal epithelial differentiation (Extended Data Fig. 7).

To investigate the translational potential of this approach, *ex vivo* expanded corneal epithelial cell sheets were cultivated as recoverable and translatable constructs, in which approximately 1.0% of cells maintained a colony-forming capability (Extended Data Fig. 8a, b). When the sheets were transplanted onto the eyes of rabbits in an experimental model of corneal epithelial stem-cell deficiency (Extended Data Fig. 8c–j), they successfully recovered a healthy corneal barrier function (Fig. 4g, Extended Data Fig. 8k) and continued to express cornea-specific proteins (Fig. 4h). This advance—that is, the generation of functional SEAM-derived ocular surface tissue with stem/progenitor cells, which can surgically repair the front of the eye—is noteworthy because, although somatic stem cells have been used to recover the ocular surface, the long-term clinical results have not been overly encouraging[18–20]. In resolving key purification steps for corneal epithelial stem/progenitor cells by applying a combination of specific antibodies to our human iPS cell-derived ocular SEAM, we are now in the position to initiate first-in-human clinical trials of anterior eye transplantation to restore visual function.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 June 2015; accepted 14 January 2016.
Published online 9 March 2016.

1. Eiraku, M. *et al.* Self-organizing optic-cup morphogenesis in three-dimensional culture. *Nature* **472,** 51–56 (2011).
2. Nakano, T. *et al.* Self-formation of optic cups and storable stratified neural retina from human ESCs. *Cell Stem Cell* **10,** 771–785 (2012).
3. Zhong, X. *et al.* Generation of three-dimensional retinal tissue with functional photoreceptors from human iPSCs. *Nature Commun.* **5,** 4047 (2014).
4. Reichman, S. *et al.* From confluent human iPS cells to self-forming neural retina and retinal pigmented epithelium. *Proc. Natl Acad. Sci. USA* **111,** 8518–8523 (2014).
5. Mellough, C. B. *et al.* IGF-1 signaling plays an important role in the formation of three-dimensional laminated neural retina and other ocular structures from human embryonic stem cells. *Stem Cells* **33,** 2416–2430 (2015).
6. Hayashi, R. *et al.* Generation of corneal epithelial cells from induced pluripotent stem cells derived from human dermal fibroblast and corneal limbal epithelium. *PLoS ONE* **7,** e45435 (2012).
7. Shalom-Feuerstein, R. *et al.* Pluripotent stem cell model reveals essential roles for miR-450b-5p and miR-184 in embryonic corneal lineage specification. *Stem Cells* **30,** 898–909 (2012).
8. Ahmad, S. *et al.* Differentiation of human embryonic stem cells into corneal epithelial-like cells by *in vitro* replication of the corneal epithelial stem cell niche. *Stem Cells* **25,** 1145–1155 (2007).
9. Brzeszczynska, J. *et al.* Differentiation and molecular profiling of human embryonic stem cell-derived corneal epithelial cells. *Int. J. Mol. Med.* **33,** 1597–1606 (2014).
10. Lavker, R. M., Tseng, S. C. & Sun, T.-T. Corneal epithelial stem cells at the limbus: looking at some old problems from a new angle. *Exp. Eye Res.* **78,** 433–446 (2004).
11. Liem, K. F. Jr, Tremml, G., Roelink, H. & Jessell, T. M. Dorsal differentiation of neural plate cells induced by BMP-mediated signals from epidermal ectoderm. *Cell* **82,** 969–979 (1995).
12. McMahon, J. A. *et al.* Noggin-mediated antagonism of BMP signaling is required for growth and patterning of the neural tube and somite. *Genes Dev.* **12,** 1438–1452 (1998).
13. Truong, T. T., Huynh, K., Nakatsu, M. N. & Deng, S. X. SSEA4 is a potential negative marker for the enrichment of human corneal epithelial stem/progenitor cells. *Invest. Ophthalmol. Vis. Sci.* **52,** 6315–6320 (2011).
14. Estey, T., Piatigorsky, J., Lassen, N. & Vasiliou, V. ALDH3A1: a corneal crystallin with diverse functions. *Exp. Eye Res.* **84,** 3–12 (2007).
15. Pearson, J. C., Lemons, D. & McGinnis, W. Modulating Hox gene functions during animal body patterning. *Nature Rev. Genet.* **6,** 893–904 (2005).
16. Mallo, M. & Alonso, C. R. The regulation of Hox gene expression during animal development. *Development* **140,** 3951–3963 (2013).
17. Krenzer, K. L. & Freddo, T. F. Cytokeratin expression in normal human bulbar conjunctiva obtained by impression cytology. *Invest. Ophthalmol. Vis. Sci.* **38,** 142–152 (1997).
18. Nishida, K. *et al.* Functional bioengineered corneal epithelial sheet grafts from corneal stem cells expanded *ex vivo* on a temperature-responsive cell culture surface. *Transplantation* **77,** 379–385 (2004).
19. Pellegrini, G. *et al.* Long-term restoration of damaged corneal surfaces with autologous cultivated corneal epithelium. *Lancet* **349,** 990–993 (1997).
20. Nakamura, T. *et al.* Transplantation of cultivated autologous oral mucosal epithelial cells in patients with severe ocular surface disorders. *Br. J. Ophthalmol.* **88,** 1280–1284 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. Baba, Y. Oie, H. Takayanagi, S. Hara, Y. Yasukawa, J. Toga and M. Yagi of Osaka University and M. Nakagawa of Kyoto University for technical assistance and scientific discussions. This work was supported in part by the project for the realization of regenerative medicine of The Japan Agency for Medical Research and Development (AMED), The Japan Science and Technology Agency (JST) and The Ministry of Health, Labour, and Welfare of Japan and the Grants-in-Aid for Scientific Research from The Ministry of Education, Culture, Sports, Science and Technology of Japan.

Author Contributions R.H., M.T. and K.N. designed the research; R.H, Y.I., R.K. and S.A. performed the *in vitro* experiments and acquired the data; Y.S., N.N., T.I. and T.S. performed animal experiments and acquired the data; K.S. provided reagents (LN511E8); R.H., Y.I. and R.K. analysed the data and wrote the respective methods and results; S.K., K.S. and A.J.Q. supervised the project; and R.H., M.T., A.J.Q. and K.N. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.N. (knishida@ophthal.med.osaka-u.ac.jp).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Human iPS cell culture. The human iPS cell lines 201B7, 253G1, and 454E2 were obtained from the RIKEN Bio Resource Center (Tsukuba, Japan)[21,22]. The 1231A3 and 1383D2 human iPS cells were provided by the Center for iPS Cell Research and Application, Kyoto University[23]. All cells were cultured in StemFit medium (Ajinomoto, Tokyo, Japan) on LN511E8-coated ($0.5\,\mu g\,cm^{-2}$) dishes[23,24]. LN511E8, produced using cGMP-banked CHO-S cells (Life Technologies, Carlsbad, CA), was obtained from Nippi (Tokyo, Japan). In part, LN511E8 was produced using human 293-F cells as previously described[12]. The 201B7 and 454E2 human iPS cell lines were used in the *in vitro* experiments, while 201B7 and 1383D2 cells were used in the animal experiments; 253G1 and 1231A3 cells were used in the supplementary experiments, the results of which are reported in Extended Data Fig. 7. All of the experiments using recombinant DNA were approved by the Recombinant DNA Committees of Osaka University and were performed according to our institutional guidelines.

Ocular cell differentiation from human iPS cells. The differentiation culture for human iPS cells was performed as indicated in Fig. 3a. First, human iPS cells were seeded on LN511E8-coated dishes at 350–700 cells cm^{-2}, after which they were cultivated in StemFit medium for 8–12 days. The culture medium was then changed to DM (differentiation medium; GMEM (Life Technologies) supplemented with 10% knockout serum replacement (KSR; Life Technologies), 1 mM sodium pyruvate (Life Technologies), 0.1 mM non-essential amino acids (Life Technologies), 2 mM L-glutamine (Life Technologies), 1% penicillin-streptomycin solution (Life Technologies) and 55 μM 2-mercaptoethanol (Life Technologies) or monothioglycerol (Wako, Osaka, Japan))[25]. In some experiments, as indicated in the Results section, Noggin (R&D systems, Minneapolis, MN), LDN-193189 (Wako) or SB-431542 (Wako) were added for the first four days. BMP4 (R&D systems) was used in some early experiments at concentrations up to 0.125 nM. This had no discernible effect on SEAM formation, however, so its use was discontinued. After four weeks of culture in DM, the medium was changed to corneal differentiation medium (CDM; DM and Cnt-20 or Cnt-PR (w/o; EGF and FGF2) (1:1, CELLnTEC Advanced Cell Systems, Bern, Switzerland) containing 5 ng ml^{-1} FGF2 (Wako), 20 ng ml^{-1} KGF (Wako) 10 μM Y-27632 (Wako) and 1% penicillin-streptomycin solution). FGF2 in CDM was not essential for corneal epithelial induction. During CDM culture (around six to eight weeks of differentiation), non-epithelial cells were removed by manual pipetting under microscopy (Extended Data Fig. 2a, b). After pipetting, the medium was changed to fresh CDM. After four weeks of culture in CDM, the medium was changed to corneal epithelium maintenance medium (CEM; DMEM/F12 (2:1), Life Technologies) containing 2% B27 supplement (Life Technologies), 1% penicillin-streptomycin solution, 20 ng ml^{-1} KGF and 10 μM Y-27632 for two to seven weeks. To achieve retinal differentiation (Fig. 2c) after four weeks of differentiation the medium was directly changed to CEM. Isolated RPE cell colonies were cultivated in CEM on separate dishes coated with LN511E8. Phase-contrast microscopic observations were performed with an Axio-observer. Z1, D1 (Carl Zeiss, Jena, Germany) and an EVOS FL Auto (Life Technologies).

Flow cytometry and cell sorting. Differentiated human iPS cells in CEM were dissociated using Accutase (Life Technologies), and resuspended in ice-cold KCM medium (DMEM without glutamine and Nutrient Mixture F-12 Ham (3:1, Life Technologies) supplemented with 5% FBS (Japan Bio Serum, Hiroshima, Japan), 0.4 μg ml^{-1} hydrocortisone succinate (Wako), 2 nM 3,3′,5-Triiodo-L-thyronine sodium salt (MP biomedicals, Santa Ana, CA), 1 nM cholera toxin (List Biological Laboratory, Campbell, CA), 2.25 μg ml^{-1} bovine transferrin HOLO form (Life Technologies), 2 mM L-glutamine, 0.5% insulin transferrin selenium solution (Life Technologies) and 1% penicillin-streptomycin solution). The harvested cells were filtered with a cell strainer (40 μm, BD Biosciences, San Diego, CA) and then stained with anti-SSEA-4 (MC813-70, Biolegend, San Diego, CA), TRA-1-60 (TRA-1-60-R, Biolegend) and CD104 (ITGB4; 58XB4, Biolegend) antibodies for 1 h on ice. After being washed twice with PBS, stained cells underwent cell sorting with a FACSAria II instrument (BD Biosciences). For intracellular protein staining, a BD Cytofix/Cytoperm (BD Biosciences) kit was used. In all of the experiments, cells were stained with non-specific isotype IgG or IgM as controls (Biolegend). The data were analysed using the BD FACSDiva Software (BD Biosciences) and the FlowJo software program (TreeStar, San Carlos, CA).

Fabrication and harvest of human iPS cell-derived corneal epithelial cell (human iCEC) Sheets. Sorted human iPS cell-derived epithelial cells obtained from zone 3 of the SEAM (human iCECs) were seeded on LN511E8 coated ($0.5\,\mu g\,cm^{-2}$) cell culture inserts or temperature-responsive dishes (UpCell, CellSeed, Tokyo, Japan) without cell passaging, and were cultured in CEM until confluence[26]. To promote maturation, the epithelial cells were cultivated in CMM (corneal epithelium maturation medium; KCM medium containing 20 ng ml^{-1} KGF and 10 μM Y-27632) for an additional 3–14 days after CEM culture. The human iCECs cultivated on temperature-responsive dishes were released from their substrate by reducing the temperature to 20 °C.

Quantitative real-time reverse-transcriptase PCR (qRT–PCR). Total RNA was obtained from differentiated human iPS cells after specific culture periods, from human epidermal keratinocytes (EKs (foreskin), Life Technologies and TaKaRa Bio, Otsu, Japan), and from human corneal limbal epithelial cells (CECs) using the RNeasy total RNA kit or the QIAzol reagent (Qiagen, Valencia, CA). Reverse transcription was performed using the SuperScript III First-Strand Synthesis System for qRT–PCR (Life Technologies) according to the manufacturer's protocol, and cDNA was used as a template for PCR. qRT–PCR was performed using the ABI Prism 7500 Fast Sequence Detection System (Life Technologies) in accordance with the manufacturer's instructions. The TaqMan MGB used in the present study are shown in Supplementary Table 2. The thermocycling program was performed with an initial cycle at 95 °C for 20 s, followed by 45 cycles at 95 °C for 3 s and 60 °C for 30 s.

Immunofluorescence staining. Research grade human skin tissue sections were obtained from US Biomax Inc. (MD, USA) and human oral mucosal tissue was obtained from Science Care (Phoenix, AZ). The cells were fixed in 4% paraformaldehyde (PFA) or cold methanol, washed with Tris-buffered saline (TBS, TaKaRa Bio) three times for 10 min and incubated with TBS containing 5% donkey serum and 0.3% Triton X-100 for 1 h to block non-specific reactions. They were then incubated with the antibodies shown in Supplementary Table 3 at 4 °C overnight or at room temperature for 3 h. The cells were again washed twice with TBS for 10 min, and were incubated with a 1:200 dilution of Alexa Fluor 488-, 568-, 647-conjugated secondary antibodies (Life Technologies) for 1 h at room temperature. Counterstaining was performed with Hoechst 33342 (Molecular Probes) before fluorescence microscopy (Axio Observer.D1, Carl Zeiss).

Haematoxylin and eosin staining. Fabricated human iCEC sheets were fixed with 10% formaldehyde neutral buffer solution (Nacalai Tesque, Kyoto, Japan). After washing with distilled water, the human iCEC sheets were embedded in paraffin from which 3-μm-thick sections were cut. These were stained with haematoxylin and eosin following deparaffinization and hydration. The sections were observed with a NanoZoomer-XR C12000 (Hamamatsu Photonics, Hamamatsu, Japan), BZ-9000 (KEYENCE, Osaka, Japan) and an Axio Observer.D1.

PAS staining. Differentiated human iPS cells (more than 12 weeks of differentiation) were fixed with 10% formaldehyde neutral buffer solution, after which PAS staining was performed with a PAS staining kit (MERCK KGaA, Darmstadt Germany) according to the manufacturer's protocol. The sections were observed with an Axio Observer.D1.

Colony-formation assay (CFA). Epithelial cells were seeded onto MMC-treated NIH-3T3 feeder layers at a density of 3,000–20,000 cells per well. These were cultivated in CMM for 7–14 days. The colonies were fixed with 10% formaldehyde neutral buffer solution and then stained with rhodamine B (Wako). Colony formation was then assessed using a dissecting microscope and the colony-forming efficiency (CFE) was calculated. For the holoclone analysis, a single human iCEC colony derived from the SEAM was cultivated on 3T3-J2 (provided by H. Green, Harvard Medical School, Boston, MA) in CMM for 7–11 days was picked up under a dissecting microscope and dissociated by TrypLE Select (Life Technologies). The dissociated human iCECs were again seeded on a MMC-treated 3T3-J2 feeder layer and cultivated in CMM for 10–13 days. The colonies were scored under a microscope and classified as holoclones, paraclones or meloclones based on previously reported methods[27].

Microarray analysis. Human CECs were harvested from corneoscleral rims (Northwest Lions Eye Bank, Seattle, WA) as reported previously[28]. Human CECs and human oral keratinocytes (OKs; ScienCell, Carlsbad, CA) along with SEAM-derived human iCECs were cultivated on LN511E8 coated cell culture inserts in CEM until confluent. They were then cultivated in CMM. Human dermal fibroblasts (DFs; ScienCell) were cultivated in DMEM/F12 (2:1) containing 10% FBS. Total RNA was obtained from human iPS cells, iCECs, CECs, OKs, DFs, and six-week differentiated iPS cells (that is, OSE) using the QIAzol reagent. A microarray analysis using Sure Print G3 human 8x60K slides (Agilent technologies, Palo Alto, CA) was performed at Takara Bio. The data were analysed using the GeneSpring GX software program (Agilent technologies). Microarray data used in this study are deposited in Gene Expression Omnibus under accession number GSE73971.

Scanning electron microscopy (SEM). The cultivated epithelial cell sheets were fixed in 2.5% glutaraldehyde (Nacalai Tesque) at 4 °C overnight. Subsequently, the sheets were washed in buffer, dehydrated with ethanol and tert-butyl alcohol (Wako), and critical point dried (JFD-320, JEOL, Tokyo, Japan). After sputter-coating with platinum in an auto fine coater (JFCL-1600, JEOL), the samples were observed by scanning electron microscopy (JSM-6510LA, JEOL) at 5 kV.

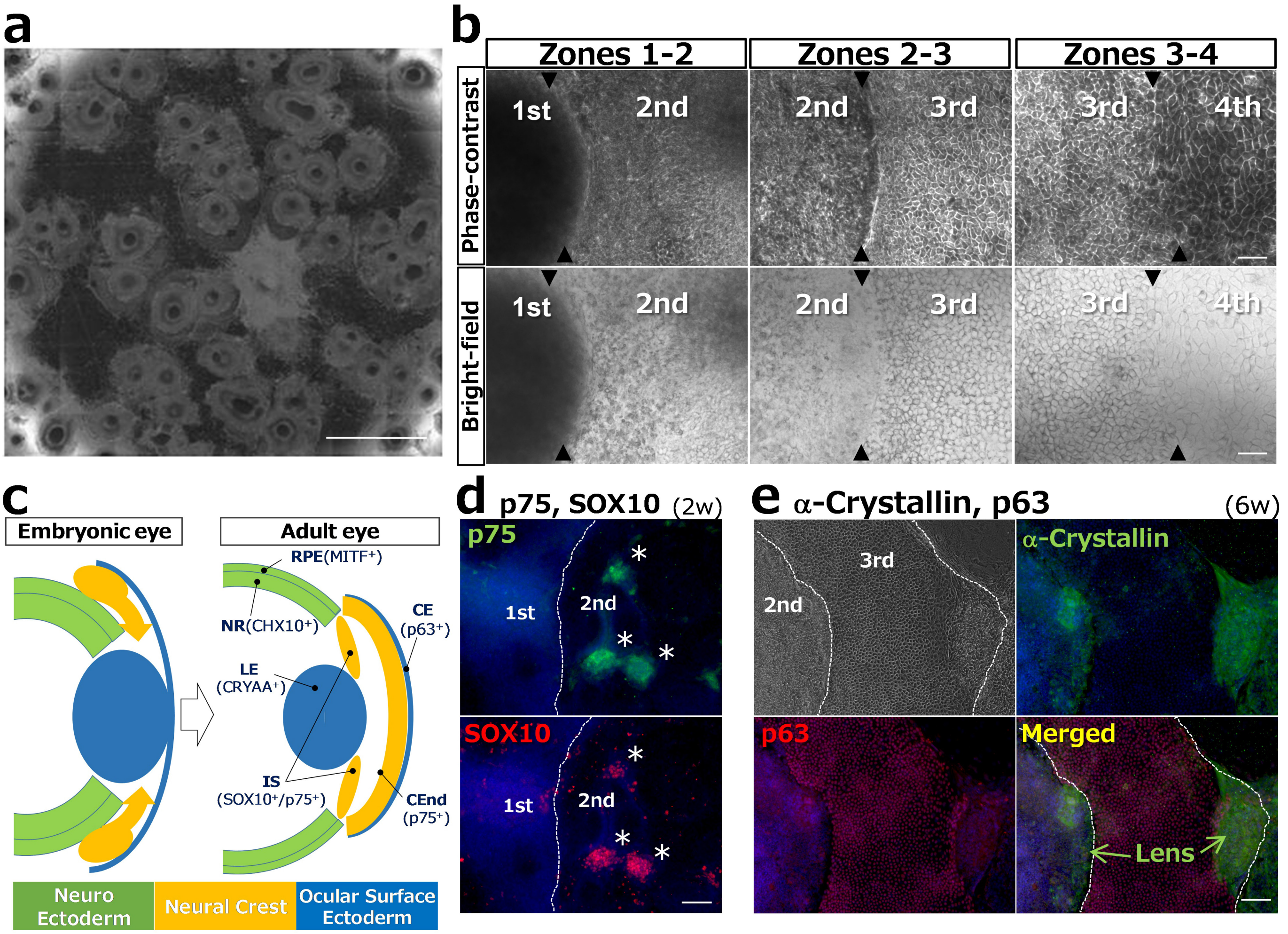
Serial cell passaging and karyotype analysis. FACS-isolated human iCECs were cultivated on MMC-treated NIH-3T3 feeder layers in CMM up to 70–80% confluence. The human iCECs were harvested using TrypLE Select following the removal of feeder cells by manual pipetting. The total cell numbers were counted, after which the cells were passaged at a 1:8 ratio onto newly prepared feeder layers. These were cultivated in CMM until sub-confluence was reached again. The G-band karyotype analysis for human iCECs was performed at Nihon Gene Research Laboratories (Sendai, Japan).

Animal experiments. All animal experimentation was performed in accordance with the ARVO Statement for the Use of Animals in Ophthalmic and Vision Research, and was approved by the animal ethics committees of Osaka University. To examine embryonic mouse eyes, pregnant females (C57/BL6, E9.5–18.5) were acquired from SLC Japan (Shizuoka, Japan). For the transplantation experiments, Female New Zealand white rabbits (2.5–3.0 kg (approximately 12–14 weeks)) were obtained from Kitayama Labes (Nagano, Japan). Harvested human iCEC sheets were grafted onto rabbit corneas, in which a total epithelial limbal stem-cell deficiency had been created following a corneal and limbal lamellar keratectomy (Extended Data Fig. 8c–j). After surgery, 0.3% ofloxacin ointment (Santen Pharmaceutical, Osaka, Japan), 0.1% betamethasone phosphate eye drops (Shionogi Pharmaceutical, Osaka, Japan) and 0.1% sodium hyaluronate eye drops (Santen Pharmaceutical) were applied three to four times per day. Triamcinolone acetonide (8 mg; Bristol Myers Squibb, Tokyo, Japan) was also administered by subconjunctival injection. Tacrolimus (0.05 $mg\,kg^{-1}$ per day, Astellas Pharma, Tokyo, Japan) and Mizoribine (4.0 $mg\,kg^{-1}$ per day Sawai Pharmaceutical, Osaka, Japan) were systemically administered using an osmotic pump (DURECT, Cupertino, CA). The corneal barrier function following surgery was assessed by 0.5% fluorescein eye drop instillation at day 7 and day 14 after surgery and the fluorescein negative area was calculated using the AxioVision software program (Carl Zeiss). Throughout the healing period, the cornea was observed with a digital slit-lamp camera (SL-7F, TOPCON, Tokyo, Japan) and 3D OCT1000 MARK II (TOPCON) or CASIA SS-1000 (TOMEY, Nagoya, Japan) machines. If an infection was found or if unexpected weight loss occurred, animals were excluded from the analysis. The rabbits were euthanized by the intravenous administration of sodium pentobarbitone 14 days after transplantation, after which the eyes were immediately enucleated for the histological analyses. No blinding or randomization was conducted to allocate animals to each group.

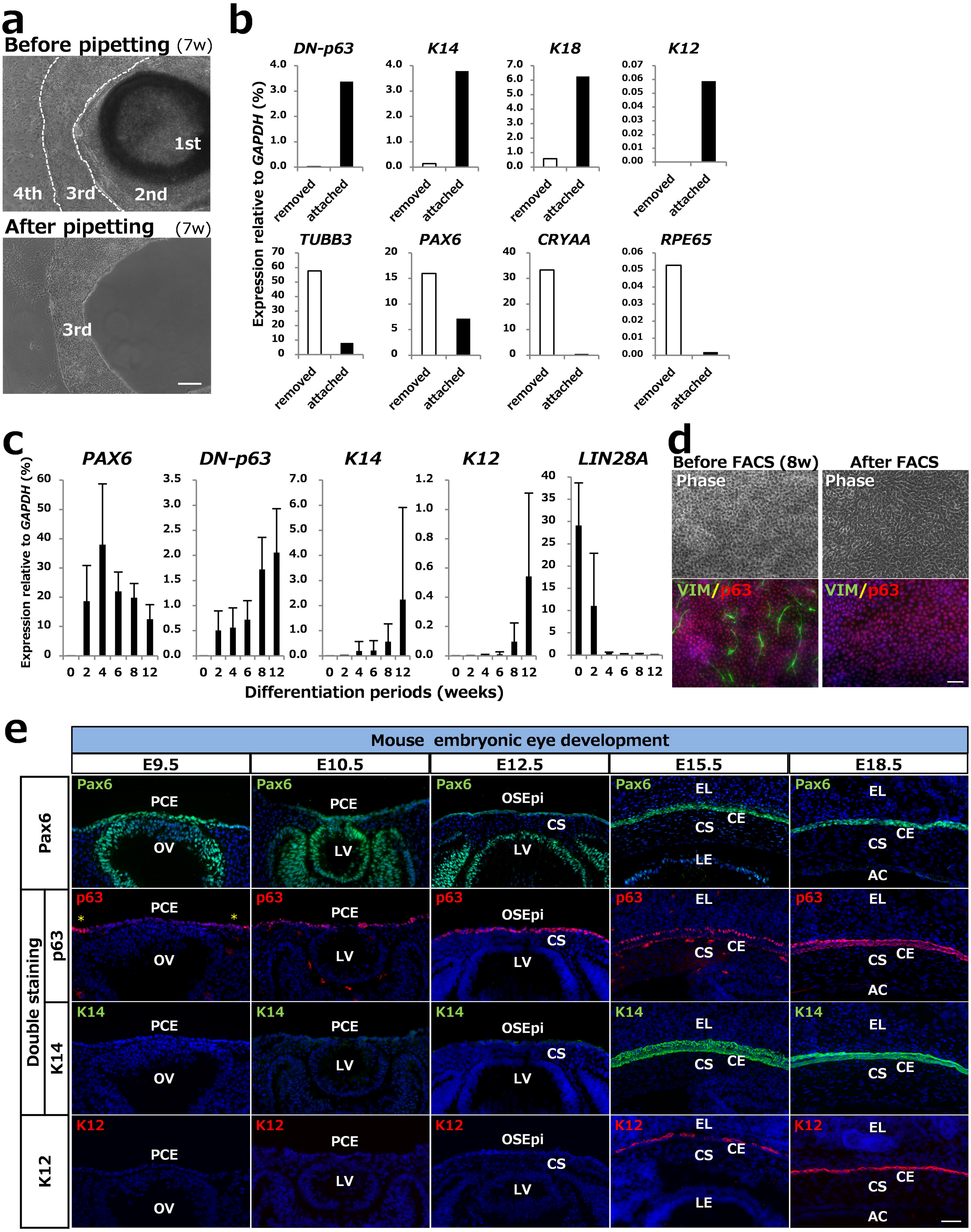
Statistical analyses. The data are expressed as means ± standard deviation (s.d.). The statistical analyses were performed using the Mann–Whitney rank sum test or Steel's test. Bonferroni's correction was applied to the data in animal experiments. All of the statistical analyses were performed using the JMP software program (SAS institute Inc., Cary, NC). No statistical methods were used to predetermine sample size.

Comprehensive technical details can be found in *Protocols Exchange*, http://dx.doi.org/10.1038/protex.2016.009.

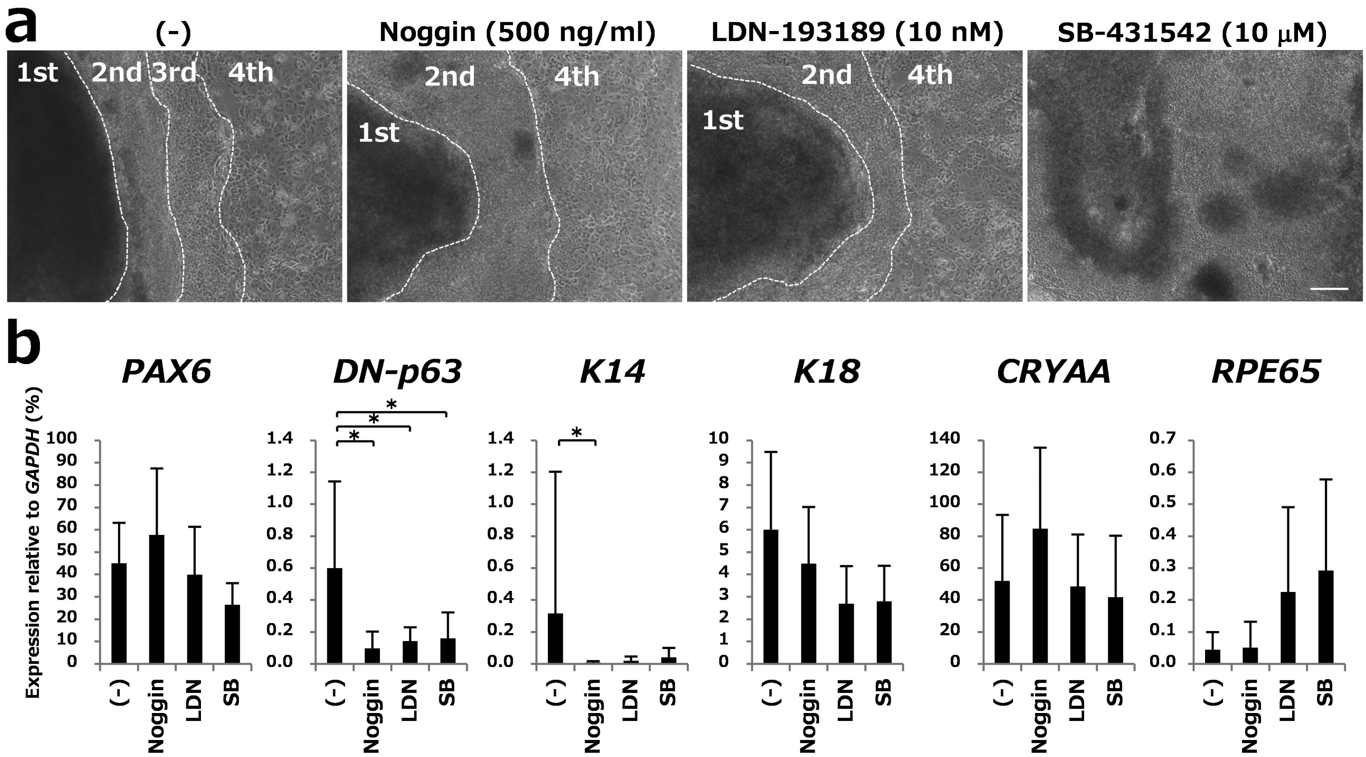
21. Nakagawa, M. *et al.* Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nature Biotechnol.* **26,** 101–106 (2008).
22. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131,** 861–872 (2007).
23. Nakagawa, M. *et al.* A novel efficient feeder-free culture system for the derivation of human induced pluripotent stem cells. *Sci. Rep.* **4,** 3594 (2014).
24. Miyazaki, T. *et al.* Laminin E8 fragments support efficient adhesion and expansion of dissociated human pluripotent stem cells. *Nature Commun.* **3,** 1236 (2012).
25. Kawasaki, H. *et al.* Generation of dopaminergic neurons and pigmented epithelia from primate ES cells by stromal cell-derived inducing activity. *Proc. Natl Acad. Sci. USA* **99,** 1580–1585 (2002).
26. Miyashita, H. *et al.* Long-term maintenance of limbal epithelial progenitor cells using Rho kinase inhibitor and keratinocyte growth factor. *Stem Cells Transl. Med.* **2,** 758–765 (2013).
27. Barrandon, Y. & Green, H. Three clonal types of keratinocyte with different capacities for multiplication. *Proc. Natl Acad. Sci. USA* **84,** 2302–2306 (1987).
28. Hayashi, R. *et al.* N-Cadherin is expressed by putative stem/progenitor cells and melanocytes in the human limbal epithelial stem cell niche. *Stem Cells* **25,** 289–296 (2007).



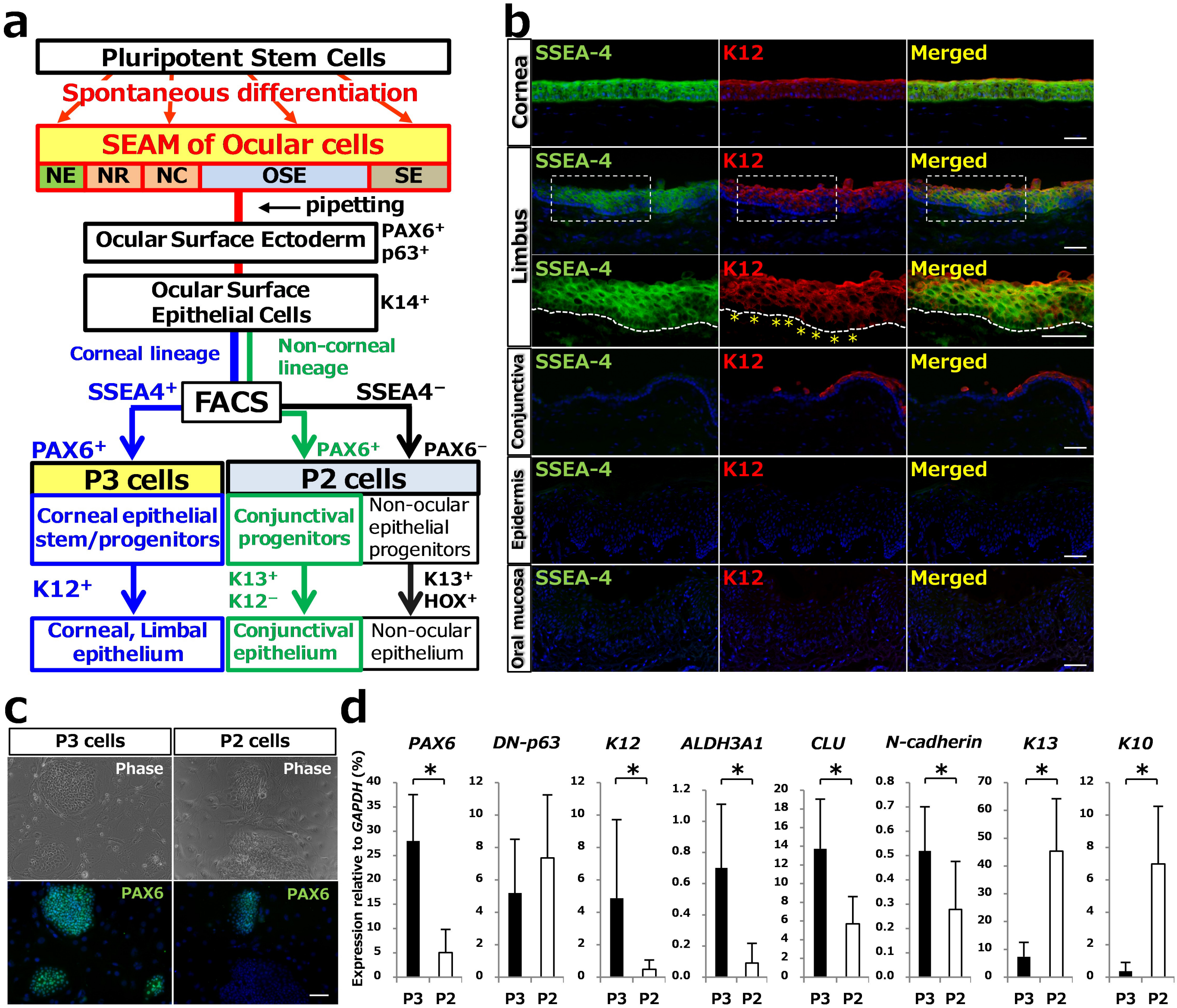
Extended Data Figure 1 | Differentiation of multiple ocular cells in the SEAM. a, A differentiated human iPS cell colony after 40 days of culture (macro photograph, representative of six independent experiments). Scale bar, 5 mm. **b**, Magnified views of each SEAM zone margin after six weeks in culture (phase-contrast and bright-field views, each representative of three independent experiments). Arrow heads indicate borders between each zone. Scale bars, 50 μm. **c**, Schematic for the development of the anterior eye. CE, corneal epithelium; CEnd, corneal endothelium; IS, iris stroma; NR, neuroretina; RPE, retinal pigment epithelium; LE, lens. **d**, Immunostaining for p75 (green) and SOX10 (red) in SEAM zones 1 and 2, two weeks (w) after the start of the differentiation culture (representative of three independent experiments). Asterisks indicate $SOX10^{+}/p75^{+}$ neural crest cells. Nuclei, blue. Scale bar, 100 μm. **e**, Immunostaining of lens differentiation marker α-crystallin (green) and epithelial marker p63 (red) in zones 2 and 3 of the SEAM after six weeks of culture (representative of three independent experiments). Nuclei, blue. Scale bar, 100 μm.



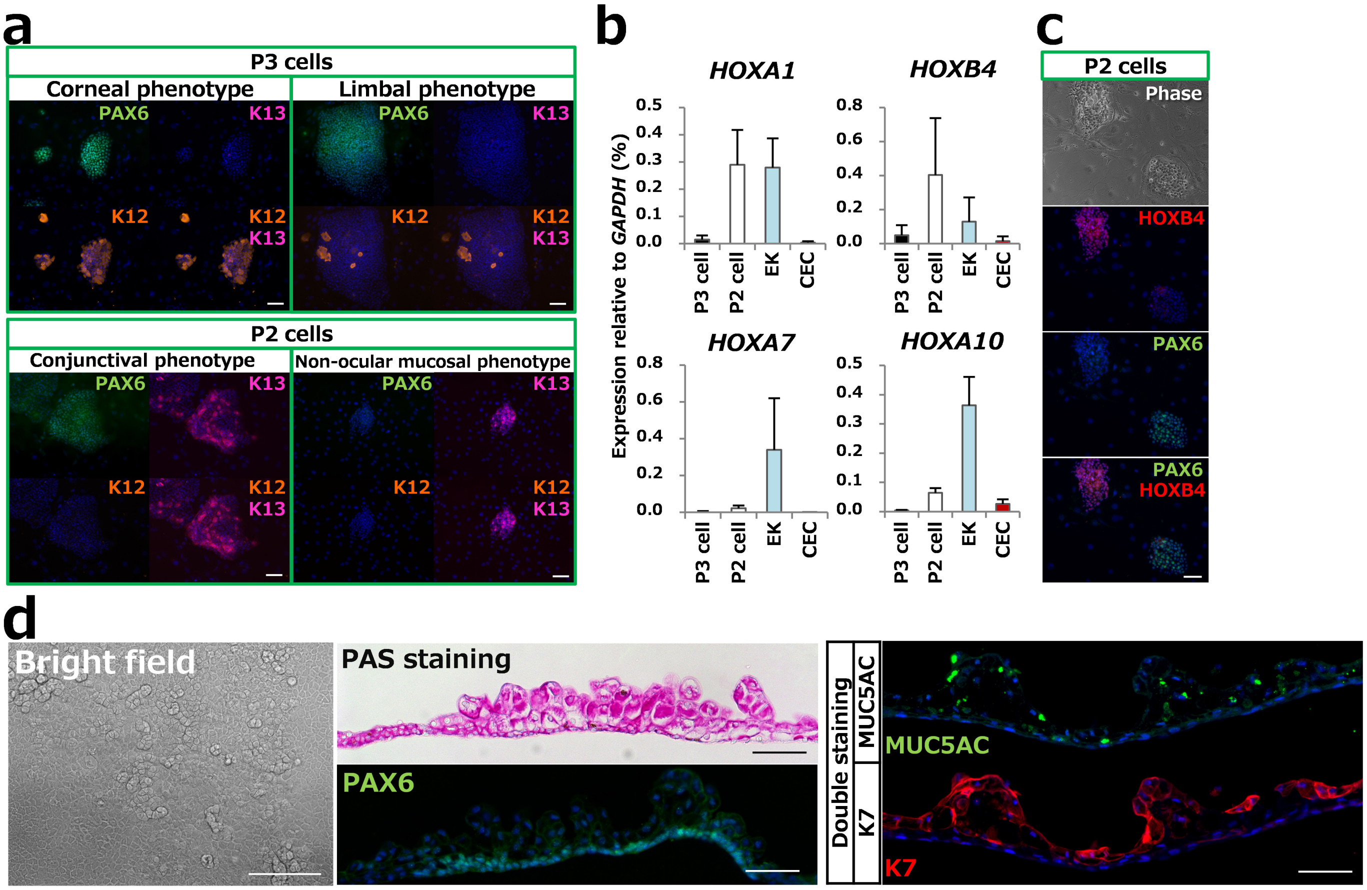
Extended Data Figure 2 | Enrichment and differentiation of a SEAM-derived ocular surface epithelium. a, The enrichment of zone 3 ocular surface epithelium by manual pipetting. Cells in the SEAM before (upper panel) and after (lower panel) pipetting are shown ($n = 1$). Scale bar, 200 μm. **b**, The expression of ocular-cell-related genes in removed and attached cells after pipetting ($n = 1$). **c**, The time course of *PAX6*, *DN-p63*, *K14*, *K12* and *LIN28A* expression during human iPS cell differentiation culture (0, 2, 4, 6, 8 and 12 weeks, each $n = 5$ independent experiments). Error bars are s.d. **d**, Immunostaining for vimentin (VIM, green) and p63 (red) before and after FACS purification conducted at 11 weeks of culture (representative of three independent experiments). VIM^{+} stroma-like cells of zone 3 were removed by FACS. Nuclei, blue. Scale bar, 50 μm. **e**, Immunostaining for anterior eye development-related markers Pax6 (green), p63 (red), K14 (green) and K12 (red) during mouse eye development (E9.5–18.5, each representative of three experiments). Nuclei, blue. Asterisks indicate p63-expressing cells. PCE, presumptive corneal epithelium; OSEpi, ocular surface epithelium; CE, corneal epithelium; CS, corneal stroma; LV, lens vesicle; LE, lens; AC, anterior chamber; EL, eyelid; OV, optic vesicle. Scale bar, 50 μm.



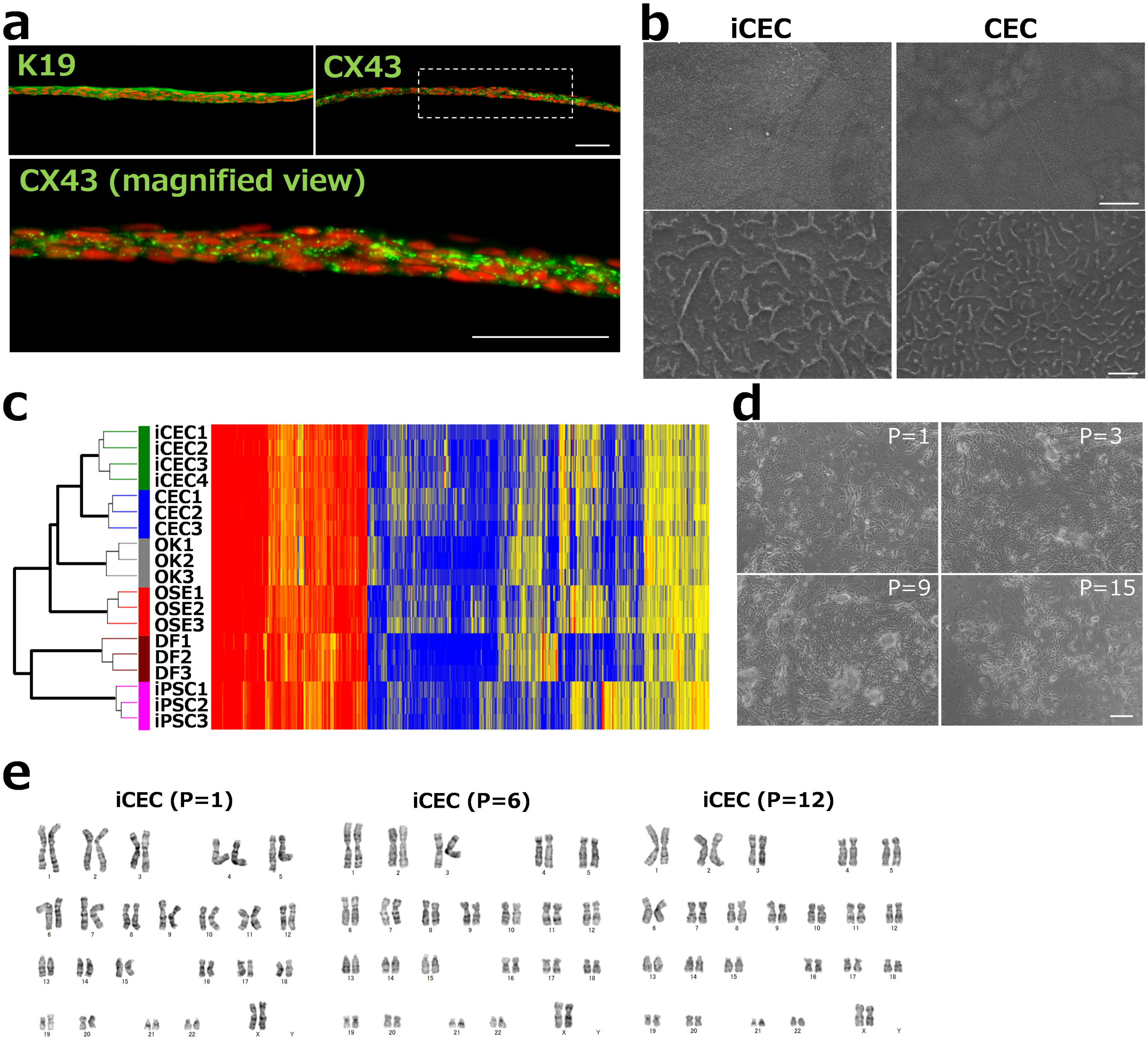
Extended Data Figure 3 | The effect of BMP/TGFβ inhibitors on the development of ocular surface epithelium in the SEAM. a, Microscopic observation of the SEAM pre-treated with Noggin, LDN-193189 (LDN) or SB-431542 (SB) for four days at the start of differentiation culture (control (−), Noggin and LDN data are representative of six independent experiments, while SB data are representative of four independent experiments). Both BMP and TGFβ inhibitors resulted in the abolishment of zone 3. Scale bar, 200 μm. **b**, The effects of BMP or TGFβ inhibitors on the expression of ocular ectoderm-related genes at five to six weeks of differentiation culture (*$P < 0.05$, control; $n = 13$: Noggin, LDN and SB; $n = 8$ independent experiments, respectively, Mann–Whitney test). Error bars are s.d.



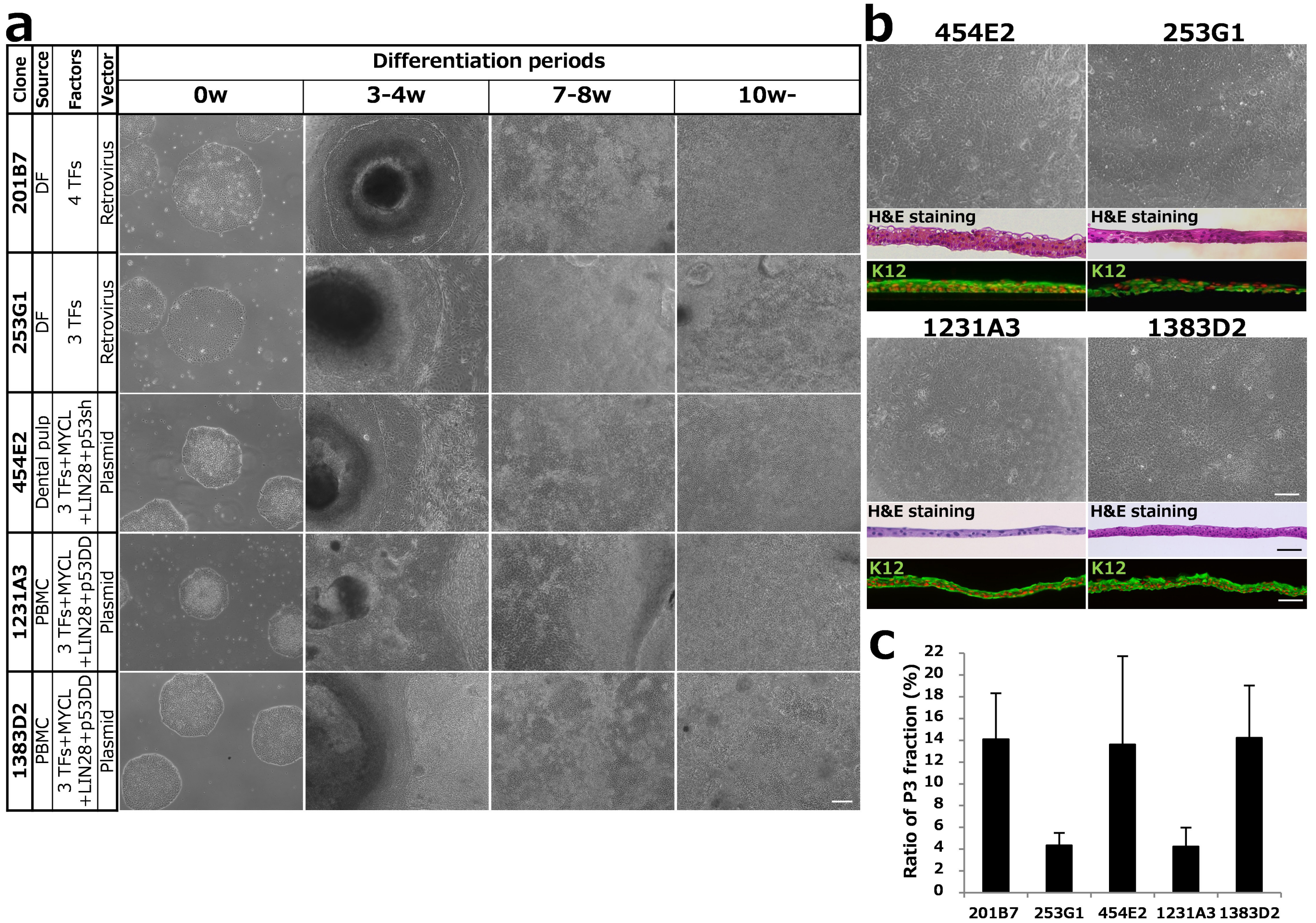
Extended Data Figure 4 | Isolation of corneal epithelial cells from the SEAM. a, Schematic for the induction of ocular surface epithelium from human iPS cells. The ocular surface ectoderm expressing PAX6 and p63 in zone 3 of the SEAM further differentiated into functional ocular surface epithelial stem cells that expressed K14. Among the ocular surface epithelial lineage cells, corneal epithelial progenitor cells were isolated by FACS as SSEA-4^{+}/ITGB4^{+} cells. Conjunctival epithelial cells (PAX6^{+}/K13^{+}/K12^{-}) were obtained as SSEA-4^{-} cells. NE, neuroectoderm; NR, neuroretina; NC, neural crest; OSE, ocular surface ectoderm; SE, surface ectoderm. **b**, Immunostaining for SSEA-4 (green) and K12 (red) in stratified epithelial tissues, including human ocular surface epithelium (cornea, limbus and conjunctiva), epidermis and oral mucosa. A magnified view of the limbus (highlighted box) indicated that SSEA-4 was expressed in all layers of the limbal epithelium, including the basal layer, which had no K12 (asterisks). No other stratified epithelium expressed SSEA-4 and K12. Nuclei, blue. Scale bars, 50 μm. (Data for corneal, limbal and conjunctival tissue; $n = 3$; epidermis and oral mucosa; $n = 1$). **c**, Immunostaining for PAX6 (green) in colonies derived from P3 and P2 cells (representative of five independent cell-sorting experiments). Nuclei, blue. Scale bar, 100 μm. **d**, The expression of corneal epithelial-specific genes and non-corneal epithelial genes in the sorted P3 cells (that is, human iPS cell-derived corneal epithelial cells, iCECs) and P2 cells. *$P < 0.05$ ($n = 7$ independent cell sorting experiments, Mann–Whitney test). Error bars are s.d.



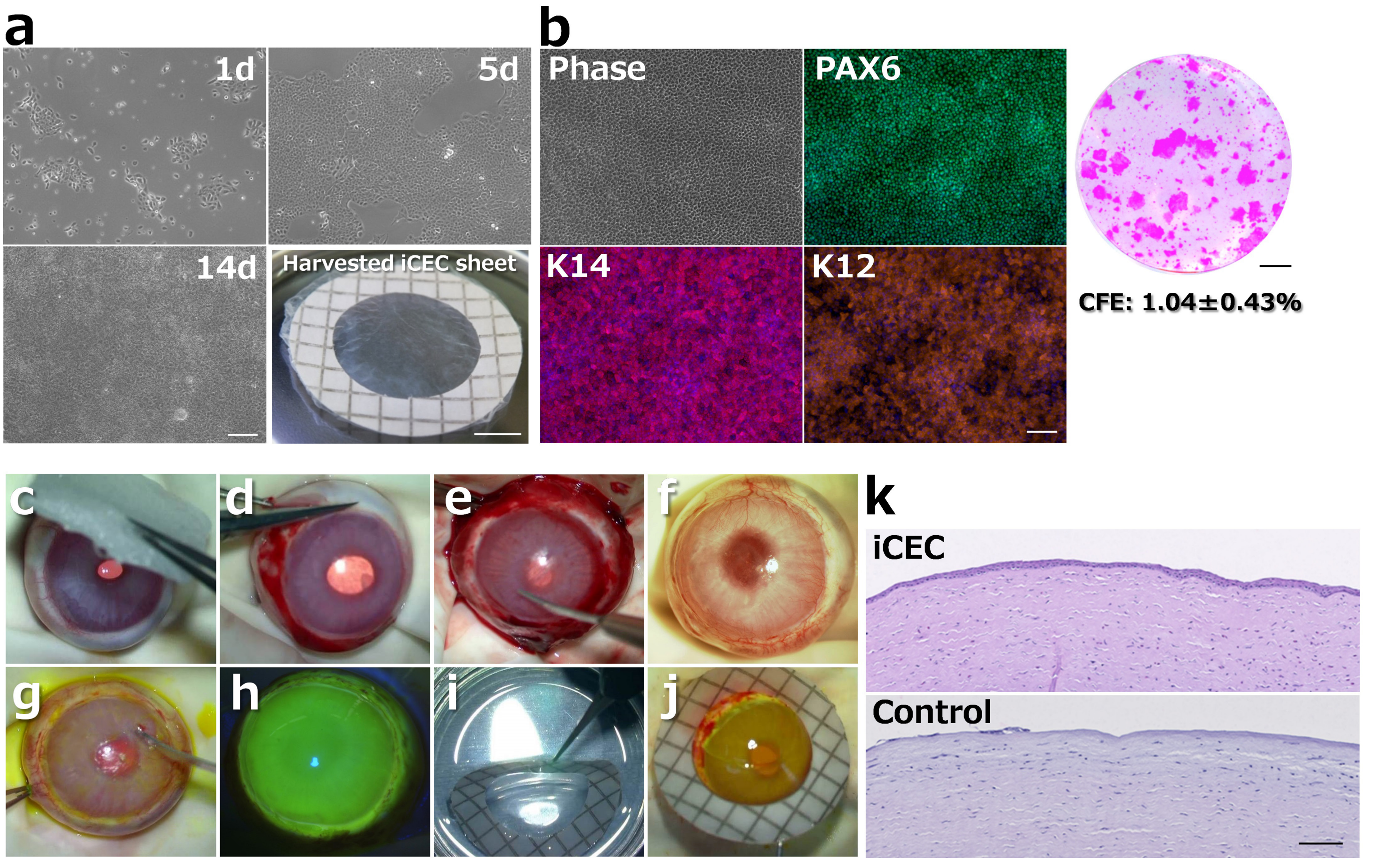
Extended Data Figure 5 | Characterization of SEAM-derived ocular surface epithelial cells. a, Triple colour immunostaining for PAX6 (green), K12 (orange) and K13 (magenta) in the epithelial colonies from sorted P3 cells (that is, SEAM-derived human iCECs) and P2 cells (representative of five independent cell-sorting experiments). Nuclei, blue. Scale bars, 100 μm. **b**, *HOX* gene expression in the sorted P3 cells, P2 cells, human epidermal keratinocytes (EKs) and human corneal limbal epithelial cells (CECs). (P3 and P2 cells; $n = 7$: human EKs and human CECs; $n = 5$ independent experiments). Error bars are s.d. **c**, The PAX6 (green) and HOXB4 (red) expression in the colonies of P2 cells ($n = 1$). Nuclei, blue. Scale bar, 100 μm. **d**, Goblet-cell-like differentiation in the SEAM-derived epithelium after long-term culture (more than 12 weeks of differentiation) without FACS in CEM ($n = 1$). Goblet-cell-like morphology was observed in presumptive P2 cell regions (left panel). These cells were PAS-positive and expressed the goblet cell markers MUC5AC (green) and K7 (red) in the superficial region and PAX6 in the basal region. Nuclei, blue. Scale bars, 50 μm.



Extended Data Figure 6 | Characterization of the SEAM-derived corneal epithelium. a, Immunostaining for K19 and CX43 (green) in the stratified SEAM-derived human iCECs (representative of $n = 3$ independent experiments). Magnified view of the dotted area is shown in the lower panel. Nuclei, red. Scale bars, 50 μm. **b**, Scanning electron microscopy of the apical surface of the stratified human iCECs (representative of two human iCEC sheets) and human CECs ($n = 1$). Scale bars, 10 μm (upper panels) and 1 μm (lower panels). **c**, Results of a hierarchical cluster analysis based on the global gene expression as examined by microarrays. Data are shown for human iPS cells ($n = 3$ technical replicates), human iPS cell-derived ocular surface ectoderm (OSE; that is, human iPS cell-derived cells after six weeks of differentiation, $n = 3$ technical replicates), human iCECs ($n = 4$ independent experiments), human oral keratinocytes (OKs, $n = 3$ technical replicates), human dermal fibroblasts (DFs, $n = 3$ technical replicates) and human corneal epithelial cells obtained from the limbus (CECs, $n = 3$ independent experiments). A total of 25,262 significantly changed genes (fold change >2.0, false discovery rate <0.05) were analysed. **d**, SEAM-derived human iCECs at passage (P) 1, 3, 9 and 15 during serial passages (representative of three independent experiments). Scale bar, 200 μm. **e**, The G-band karyotype of human iCECs at $P = 1$, 6 and 12 ($n = 1$, respectively).



Extended Data Figure 7 | Induction of corneal epithelial cells from different human iPS cell clones. a, SEAM formation patterns of five different human iPS cell clones (201B7 and 1383D2; representative of three independent experiments: 253G1, 454E2 and 1231A3; $n = 1$). PBMC, peripheral blood mononuclear cell; TF, transcription factor. Scale bar, 200 µm. **b**, Stratified human iCECs from 454E2, 253G1, 1231A3 and 1383D2 clones *in vitro* (454E2, 253G1 and 1231A3, $n = 1$; 1383D2, representative of three independent experiments). Phase-contrast microscopy (upper panel), haematoxylin and eosin (H&E) staining (middle panel) and immunostaining for K12 (lower panel, green) are shown. Nuclei, red. Scale bars, 200 µm (phase-contrast), 50 µm (H&E staining and immunostaining). **c**, Efficiency of corneal epithelial differentiation of the various human iPS cell clones (201B7, $n = 23$; 253G1, $n = 5$; 454E2, $n = 11$; 1231A3, $n = 9$ and 1383D2, $n = 6$ independent experiments). Error bars are s.d.



Extended Data Figure 8 | Transplantation of the human iCEC sheet to repair the ocular surface. a, Phase-contrast microscopy of SEAM-derived human iCECs cultivated on a temperature-responsive dish at 1, 5 and 14 days (representative of three independent experiments). A harvested human iCEC sheet is shown in the lower right panel. Scale bar, 200 μm (phase-contrast), 5 mm (macro photo). **b**, Immunostaining for PAX6 (green), K12 (orange) and K14 (magenta) in the human iCEC sheets after 21 days in culture (left panels, representative of three human iCEC sheets). The right panel shows an image used for a colony-forming assay (CFA) for the human iCEC sheets (15,000 cells per well, representative of eight independent experiments). The colony-forming efficiency (CFE) was 1.04 ± 0.43% (s.d., $n = 8$ independent experiments). Nuclei, blue. Scale bars, 100 μm (immunostaining) and 5 mm (CFA). **c**, Treatment of rabbit ocular surface with a surgical swab soaked in 99% ethanol for 30–60 s. **d**, **e**, Elimination of corneal epithelial stem cells by the surgical removal of corneal and limbal epithelial tissue (that is, lamellar keratectomy). **f**, The ocular surface was invaded by conjunctival tissue and vessels at postoperative day 28. **g**, Removal of the conjunctival tissue that covered the ocular surface. **h**, Widespread fluorescein staining of the ocular surface after surgical removal of epithelial tissue. **i**, The harvest of a human iCEC sheet from a temperature-responsive dish after lowering the temperature. **j**, Transplantation of the human iCEC sheet onto the rabbit cornea with a surgically induced corneal epithelial stem-cell deficiency treated as described in **c**–**h**. **k**, H&E staining for human iCEC-sheet-transplanted and control corneas on postoperative day 14 ($n = 6$ animal transplantation experiments). Scale bar, 100 μm.

doi:10.1038/nature17180

Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys

Travis K. Warren[1,2], Robert Jordan[3], Michael K. Lo[4], Adrian S. Ray[3], Richard L. Mackman[3], Veronica Soloveva[1,2], Dustin Siegel[3], Michel Perron[3], Roy Bannister[3], Hon C. Hui[3], Nate Larson[3], Robert Strickley[3], Jay Wells[1], Kelly S. Stuthman[1], Sean A. Van Tongeren[1], Nicole L. Garza[1], Ginger Donnelly[1], Amy C. Shurtleff[1], Cary J. Retterer[1], Dima Gharaibeh[1], Rouzbeh Zamani[1], Tara Kenny[1], Brett P. Eaton[1], Elizabeth Grimes[1], Lisa S. Welch[1]†, Laura Gomba[1,2], Catherine L. Wilhelmsen[1], Donald K. Nichols[1], Jonathan E. Nuss[1,2], Elyse R. Nagle[1], Jeffrey R. Kugelman[1], Gustavo Palacios[1], Edward Doerffler[3], Sean Neville[3], Ernest Carra[3], Michael O. Clarke[3], Lijun Zhang[3], Willard Lew[3], Bruce Ross[3], Queenie Wang[3], Kwon Chun[3], Lydia Wolfe[3], Darius Babusis[3], Yeojin Park[3], Kirsten M. Stray[3], Iva Trancheva[3], Joy Y. Feng[3], Ona Barauskas[3], Yili Xu[3], Pamela Wong[3], Molly R. Braun[5], Mike Flint[4], Laura K. McMullan[4], Shan-Shan Chen[3], Rachel Fearns[5], Swami Swaminathan[3], Douglas L. Mayers[1]†, Christina F. Spiropoulou[4], William A. Lee[3], Stuart T. Nichol[4], Tomas Cihlar[3] & Sina Bavari[1,2]

The most recent Ebola virus outbreak in West Africa, which was unprecedented in the number of cases and fatalities, geographic distribution, and number of nations affected, highlights the need for safe, effective, and readily available antiviral agents for treatment and prevention of acute Ebola virus (EBOV) disease (EVD) or sequelae[1]. No antiviral therapeutics have yet received regulatory approval or demonstrated clinical efficacy. Here we report the discovery of a novel small molecule GS-5734, a monophosphoramidate prodrug of an adenosine analogue, with antiviral activity against EBOV. GS-5734 exhibits antiviral activity against multiple variants of EBOV and other filoviruses in cell-based assays. The pharmacologically active nucleoside triphosphate (NTP) is efficiently formed in multiple human cell types incubated with GS-5734 *in vitro*, and the NTP acts as an alternative substrate and RNA-chain terminator in primer-extension assays using a surrogate respiratory syncytial virus RNA polymerase. Intravenous administration of GS-5734 to nonhuman primates resulted in persistent NTP levels in peripheral blood mononuclear cells (half-life, 14 h) and distribution to sanctuary sites for viral replication including testes, eyes, and brain. In a rhesus monkey model of EVD, once-daily intravenous administration of 10 mg kg^{-1} GS-5734 for 12 days resulted in profound suppression of EBOV replication and protected 100% of EBOV-infected animals against lethal disease, ameliorating clinical disease signs and pathophysiological markers, even when treatments were initiated three days after virus exposure when systemic viral RNA was detected in two out of six treated animals. These results show the first substantive post-exposure protection by a small-molecule antiviral compound against EBOV in nonhuman primates. The broad-spectrum antiviral activity of GS-5734 *in vitro* against other pathogenic RNA viruses, including filoviruses, arenaviruses, and coronaviruses, suggests the potential for wider medical use. GS-5734 is amenable to large-scale manufacturing, and clinical studies investigating the drug safety and pharmacokinetics are ongoing.

The 2013–2016 outbreak of EVD in West Africa was the largest and most complex EBOV outbreak in the recorded history of the disease, with >28,000 EVD cases and >11,000 reported deaths[1]. Medical infrastructures in Guinea, Sierra Leone, and Liberia were seriously impacted by a loss of >500 healthcare workers[1]. Additionally, EVD-related sequelae (joint and muscle pain, as well as neurological, ophthalmic, and other symptoms) together with viral persistence and recrudescence in individuals who survived the acute disease have been documented[2–5].

EBOV is a single-stranded negative-sense non-segmented RNA virus from the *Filoviridae* family. In addition to EBOV, other related viruses, namely Marburg, Sudan, and Bundibugyo viruses, have caused outbreaks with high fatality rates[6]. Although the efficacy of various experimental small molecules and biologics have been assessed in EVD animal models and in multiple clinical trials during the West African outbreak[7–18], there are no therapeutics for which clinical efficacy and safety have been established for treatment of acute EVD or its sequelae. The availability of broadly effective antiviral(s) with a favourable benefit/risk profile would address a serious unmet medical need for the treatment of EBOV infection.

A 1′-cyano-substituted adenine C-nucleoside ribose analogue (Nuc) exhibits antiviral activity against a number of RNA viruses[19]. The mechanism of action of Nuc requires intracellular anabolism to the active triphosphate metabolite (NTP), which is expected to interfere with the activity of viral RNA-dependent RNA-polymerases (RdRp). Structurally, the 1′-cyano group provides potency and selectivity towards viral RNA polymerases, but because of slow first phosphorylation kinetics, modification of parent nucleosides with monophosphate promoieties has the potential to greatly enhance intracellular NTP concentrations[20]. GS-5734, the single *Sp* isomer of the 2-ethylbutyl L-alaninate phosphoramidate prodrug (Supplementary Information), effectively bypasses the rate-limiting first phosphorylation step of the Nuc (Fig. 1a). In human monocyte-derived macrophages, incubation with GS-5734 caused rapid loading of cells with high levels of NTP that persist with a half-life ($t_{1/2}$) of 24 h following removal of GS-5734 (Extended Data Fig. 1a), resulting in up to 30-fold higher levels compared to incubation with Nuc (Fig. 1b). In cell-based assays, GS-5734 is active against a broad range of filoviruses including Marburg virus and several variants of EBOV (Fig. 1c). GS-5734 inhibits EBOV replication in multiple relevant human cell types including primary macrophages and human endothelial cells with half-maximum effective concentration (EC_{50}) values of 0.06 to 0.14 μM (Table 1). As expected, the parent Nuc was less active, with EC_{50} values of 0.77 to >20 μM. Treatment with GS-5734 of liver Huh-7 cells infected with the EBOV Makona variant, isolated during the West African outbreak, resulted in profound dose-dependent reductions in viral RNA production and infectious virus yield (Extended Data Fig. 2). GS-5734 and Nuc inhibited replication of other human RNA viral pathogens including

[1]United States Army Medical Research Institute of Infectious Diseases, Frederick, Maryland 21702, USA. [2]United States Army Medical Research Institute of Infectious Diseases, Therapeutic Development Center, Frederick, Maryland 21702, USA. [3]Gilead Sciences, Foster City, California 94404, USA. [4]Centers for Disease Control and Prevention, Atlanta, Georgia 30333, USA. [5]Boston University School of Medicine, Boston, Massachusetts 02118, USA. †Present addresses: LOKET Consulting, Clarksburg, Maryland 20871, USA (L.S.W.); Cocrystal Pharma, Tucker, Georgia 30084, USA (D.L.M.).

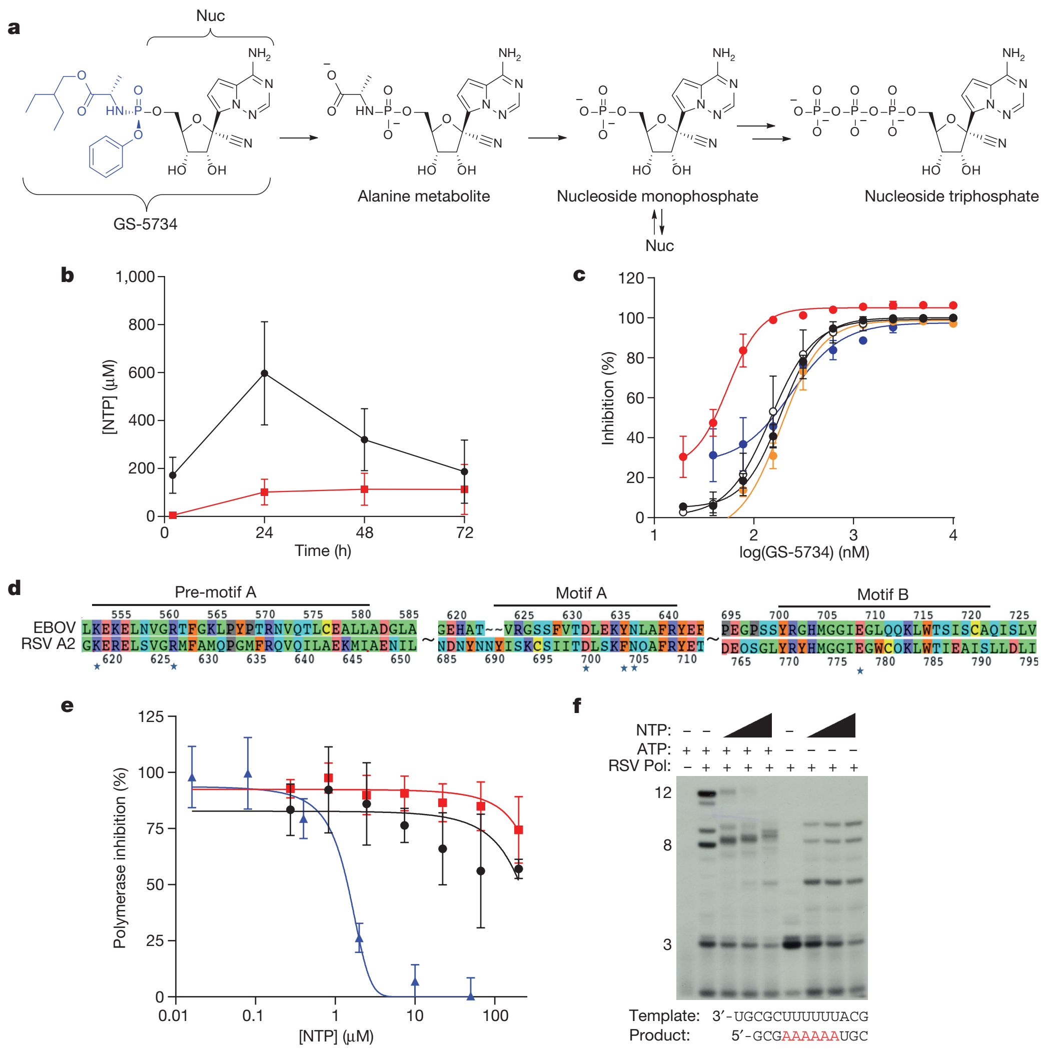


Figure 1 | Metabolism and mechanism of antiviral activity of GS-5734. **a**, Chemical structures of GS-5734 and metabolic conversion to NTP. **b**, NTP formation in human monocyte-derived macrophages following 72-h incubation with 1 μM GS-5734 (black) or Nuc (red); mean ± s.d., from 3 donors. **c**, Antiviral activity of GS-5734 in HeLa cells against EBOV Makona (black symbols), EBOV Kikwit (open symbols), Marburg (red), Bundibugyo (orange), Sudan (blue) viruses; mean ± s.d. from triplicates. **d**, Amino acid sequence homology of EBOV and RSV strain A2 RdRp active site. Asterisks indicate residues predicted to contact NTP. **e**, Inhibition of RSV RdRp (blue), but not human RNA Pol II (black) or mitochondrial RNA (red) polymerases by NTP; mean ± s.d., $n = 3$ biological replicates. **f**, NTP-induced RNA chain termination in RSV RdRp primer-extension assay. For gel source data, see Supplementary Fig. 1.

respiratory syncytial virus (RSV), Junín virus, Lassa fever virus, and Middle East respiratory syndrome virus, but was inactive against alphaviruses or retroviruses (Table 1). Previous studies have reported activity of Nuc against flaviviruses, parainfluenza virus type 3, and severe acute respiratory syndrome associated coronavirus, but little or no activity against West Nile, influenza A, or Coxsackie A viruses[19,21]. The antiviral activity of GS-5734 was selective, as demonstrated by low cytotoxicity in a wide range of human primary cells and cell lines (Extended Data Table 1).

Isolation and expression of EBOV RdRp has been elusive, but computational analysis of the catalytic palm subdomain demonstrated high sequence and structure homology with RSV RdRp[22] (Fig. 1d, Extended Data Fig. 3). Consistent with the proposed mechanism of action, NTP inhibited RSV RdRp-catalysed RNA synthesis (Fig. 1e) by incorporating into the nascent viral RNA transcript and causing its premature termination (Fig. 1f). In contrast, NTP did not inhibit human RNA polymerases (Fig. 1e). These data suggest that GS-5734 selectively inhibits EBOV replication by targeting its RdRp and inhibiting viral RNA synthesis following efficient intracellular conversion to NTP.

Rodent models were not suitable for GS-5734 *in vivo* efficacy evaluations because high serum esterase activity, present in many rodent species, degrades the GS-5734 promoiety and adversely impacts its pharmacokinetic profile[23]. Like humans, rhesus monkeys do not express high levels of serum esterase; rhesus lymphoid cells efficiently activated GS-5734 *in vitro*, although NTP levels were reduced relative to human cells (Extended Data Fig. 1b). In rhesus monkeys, intramuscular inoculation with clinically derived wild-type EBOV produces a fulminant lethal disease with pathophysiological responses that closely resemble human EVD cases[24,25], and nonhuman primates (NHP) are considered the most relevant EVD models well-suited for evaluating the efficacy of antiviral interventions when trials in infected humans are not feasible.

GS-5734 pharmacokinetics, metabolism, and distribution were examined in NHPs. Upon intravenous administration of a 10 mg kg^{-1} dose in rhesus monkeys, GS-5734 exhibited a short plasma half-life ($t_{1/2} = 0.39$ h) with fast systemic elimination followed by the sequential appearance of transient systemic levels of the key intracellular intermediate alanine metabolite and more persistent levels of Nuc (Fig. 2a). GS-5734 rapidly distributed into peripheral blood mononuclear cells (PBMCs), and efficient conversion to NTP was apparent within 2 h of dose administration. In PBMCs, NTP represents the predominant metabolite and was persistent with a $t_{1/2}$ of 14 h and levels required for >50% virus inhibition for 24 h (Fig. 2a, Extended Data Fig. 1c).

Table 1 | Antiviral activity of GS-5734 and Nuc

	Antiviral activity; EC_{50}/EC_{90} (μM)	
	GS-5734	Nuc
EBOV		
Primary macrophages*	0.086/0.18	>20/>20
HeLa cells†	0.14/0.41	>20/>20
HFF-1*	0.13/0.26	>20/>20
HMVEC-TERT cells†	0.06/0.22	0.77/3.12
Huh-7 cells†	0.07/0.22	1.49/6.04
Other human RNA viruses		
RSV‡	0.019/0.051	0.75/2.84
JUNV§	0.47/1.33	
LASV§	1.48/2.80	
MERS§	0.34/4.24	
CHIV§	>20/>20	
VEEV§	>20/>20	
HIV-1§	>20/>20	

Half-maximum cytotoxic concentration (CC_{50}) values for all compounds in primary human cells and human cell lines were greater than the highest concentration tested (>20 μM).
*Mean values from duplicated titrations conducted in differentiated macrophages or HFF-1 cells in a single experiment ($n = 1$). Cells were infected with EBOV (Makona) for antiviral activity determination.
†Mean values from quadruplicate (HMVEC-TERT) or duplicate (Huh-7) titrations generated from two experiments ($n = 2$) or from multiple experiments ($n = 6$) for HeLa cells. Cells were infected with a replication-competent reporter virus (EBOV–GFP) or wild-type EBOV strain Zaire (HeLa) for antiviral activity determination.
‡Mean values from two (GS-5734) or four (Nuc) independent experiments with each drug dilution tested in triplicate against the respiratory syncytial virus (RSV).
§Mean values from duplicate titrations with each drug concentration tested in quadruplicate from a single experiment ($n = 1$).
JUNV, Junín virus; LASV, Lassa fever virus; MERS, Middle East respiratory syndrome coronavirus; CHIV, Chikungunya virus, VEEV, Venezuelan equine encephalitis virus; HIV-1, human immunodeficiency virus type 1.

In cynomolgus monkeys, intravenous administration of a 10 mg kg^{-1} dose of [^{14}C]GS-5734 demonstrated that the drug-derived material distributed to testes, epididymis, eyes, and brain within 4 h of administration (Fig. 2b). Levels in the brain at 4 h were low relative to other tissues, but remained detectable above the drug plasma levels 168 h after dose administration. Taken together, the pharmacokinetic analysis indicates that once-daily dosing of GS-5734 provides sustained intracellular NTP levels and efficiently delivers drug metabolites to sanctuary sites where virus may persist.

To evaluate the *in vivo* efficacy of GS-5734, we conducted a sequential two-part adaptive design study in EBOV-infected rhesus monkeys (Fig. 2c). In part 1, animals intramuscularly inoculated with EBOV were administered a 12-day treatment of vehicle ($n = 3$) or 3 mg kg^{-1} GS-5734 beginning on day 0 (d0; 30–90 min following virus challenge) or day 2 (d2) ($n = 6$ per treatment group). Regardless of the time of initiation, GS-5734 treatment conferred an antiviral effect by reducing systemic viraemia relative to vehicle and an improved survival of 33% (2 out of 6) in the 3 mg kg^{-1} d0 group and 66% (4 out of 6) in the 3 mg kg^{-1} d2 group (Fig. 2d, e, Extended Data Fig. 4, Extended Data Tables 2, 3); however, mortalities observed in both treatment groups suggested that drug exposure at 3 mg kg^{-1} was suboptimal. In part 2 of the efficacy study, GS-5734 was administered once at a loading dose of 10 mg kg^{-1} followed by once-daily 3 mg kg^{-1} doses beginning either 2 days (10/3 mg kg^{-1} d2) or 3 days (10/3 mg kg^{-1} d3) after virus exposure, or 10 mg kg^{-1} doses were administered beginning 3 days after virus exposure (10 mg kg^{-1} d3; $n = 6$ per group). All 12 animals in which GS-5734 treatments were initiated 3 days after virus exposure survived to the end of the in-life phase (Fig. 2d). However, the antiviral effects were consistently greater in animals administered repeated 10 mg kg^{-1} GS-5734 doses (Fig. 2e, f, Extended Data Fig. 4, Extended Data Tables 2, 3). On day 4, plasma viral RNA was significantly decreased ($P < 0.05$), with geometric means reduced by $\geq 1.7 \log_{10}$ in all GS-5734-treated groups compared with combined vehicle-treated groups (Fig. 2e, f, Extended Data Table 3), and on days 5 and 7, when the geometric mean viral RNA concentration of the vehicle group exceeded 10^9 copies ml^{-1}, viral RNA was detected at levels less than the lower limit of quantitation (8×10^4 RNA copies ml^{-1}) in 4 of 6 animals in the 10 mg kg^{-1} d3 group. Deep sequencing analysis of the EBOV RdRp (*L*) gene from all plasma samples positive for viral RNA showed no evidence of genotypic changes potentially associated with the emergence of GS-5734-resistant

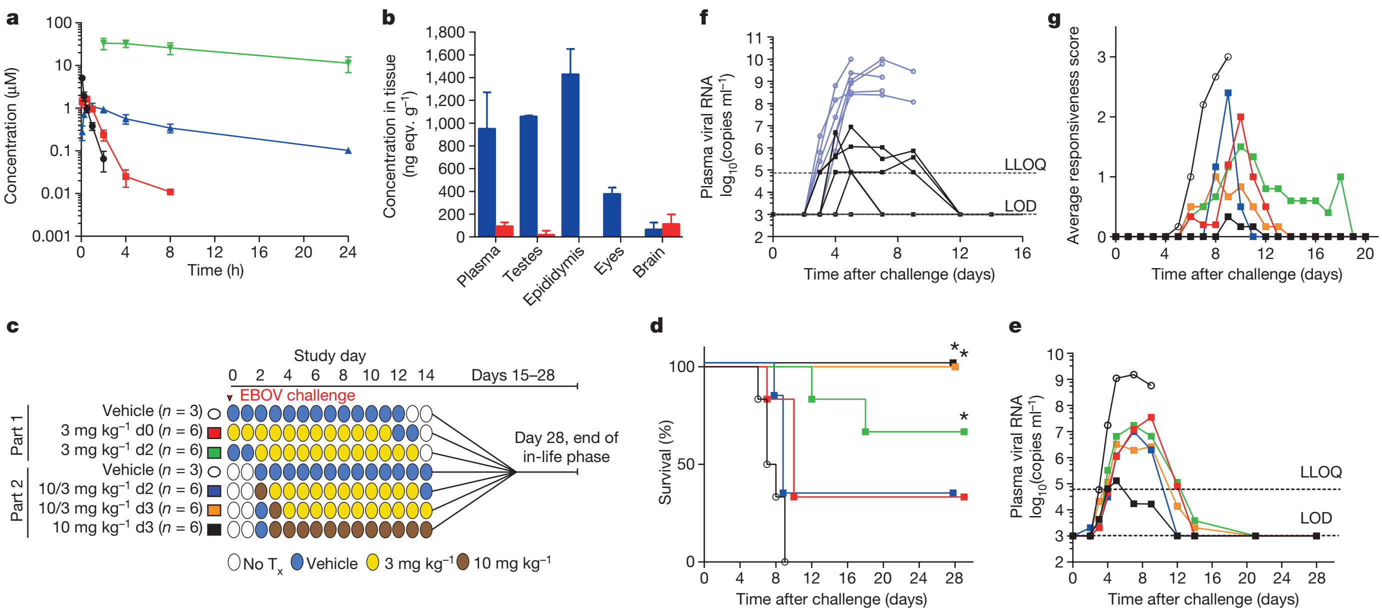


Figure 2 | GS-5734 pharmacokinetics and post-exposure protection against EBOV in rhesus monkeys. a, Pharmacokinetics following intravenous administration of 10 mg kg^{-1} GS-5734 dose in healthy rhesus macaques (mean ± s.d., $n = 3$). Plasma GS-5734 (black), alanine metabolite (red), and Nuc (blue); NTP in PBMCs (green). **b**, Tissue distribution of [^{14}C]GS-5734 and metabolites at 4 h (blue) and 168 h (red) following intravenous 10 mg kg^{-1} GS-5734 dose in healthy cynomolgus macaques (mean ± s.d., $n = 3$). **c**, Experimental design for GS-5734 efficacy evaluations in rhesus monkeys. No T$_x$, no treatment. **d**, Kaplan–Meier survival curves. *$P < 0.05$ for treatment versus vehicle groups assessed by log-rank analysis using Dunnett–Hsu procedure to adjust for multiple comparisons. **e**, Group geometric mean of plasma viral RNA concentrations; LLOQ, lower limit of quantitation; LOD, limit of detection. **f**, Individual plasma viral RNA in vehicle (blue) or 10 mg kg^{-1} GS-5734 (black) groups. **g**, Group average clinical disease score. **d**, **e**, **g**, Black (open symbols), vehicle; red, 3 mg kg^{-1} d0; green, 3 mg kg^{-1} d2; blue, 10/3 mg kg^{-1} d2; orange, 10/3 mg kg^{-1} d3; black (closed symbols), 10 mg kg^{-1} d3; $n = 6$ animals per group. Error bars omitted for clarity (**e**, **g**); *x* axes truncated to emphasize acute disease phase (**f**, **g**).

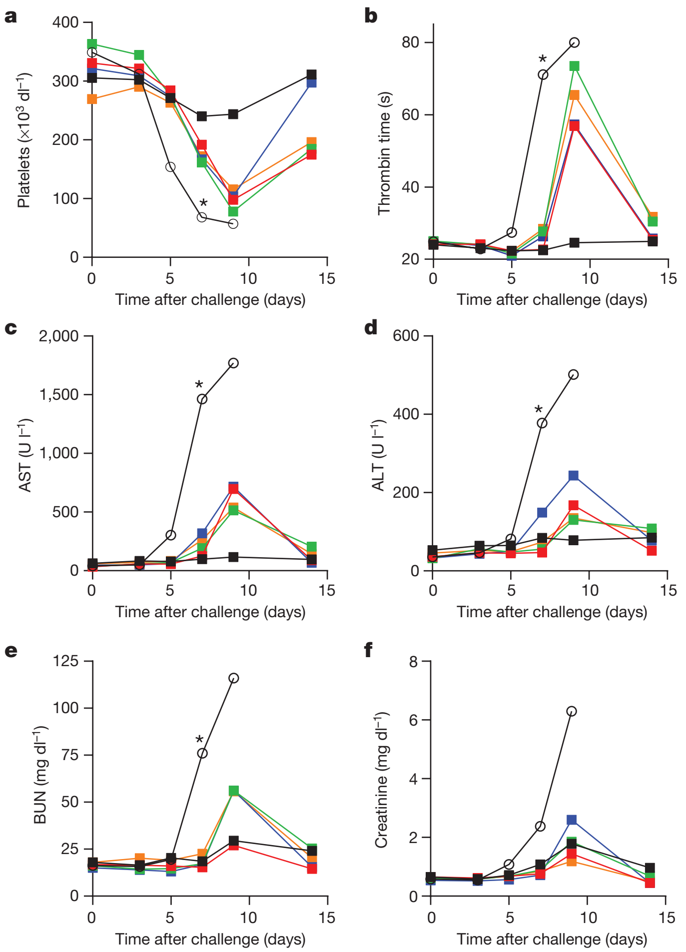


Figure 3 | Amelioration of EVD clinical pathology by GS-5734 in rhesus monkeys. a–f, Group mean ($n = 6$ per group) values of platelets (**a**), thrombin time (**b**), aspartate aminotransferase (AST, **c**), alanine aminotransferase (ALT, **d**), blood urea nitrogen (BUN, **e**), and creatinine (**f**). Black (open symbols), vehicle; red, 3 mg kg^{-1} d0; green, 3 mg kg^{-1} d2; blue, 10/3 mg kg^{-1} d2; orange, 10/3 mg kg^{-1} d3; black (closed symbols), 10 mg kg^{-1} d3. Error bars omitted for clarity; x axes truncated at day 15. *$P < 0.05$ for comparison of mean change from day 0 of vehicle and 10 mg kg^{-1} d3 groups at day 7 using Wilcoxon rank-sum test without adjustment for multiple comparisons.

EBOV variants (Extended Data Table 5). The 10 mg kg^{-1} d3 GS-5734 regimen was associated with amelioration of EVD-related clinical disease signs (Fig. 2g, Extended Data Fig. 4) and markers of coagulopathy and end-organ pathophysiology (Fig. 3a–f, Extended Data Table 4, Extended Data Fig. 5). Although greater survival was observed in the 10/3 mg kg^{-1} d3 group than the 10/3 mg kg^{-1} d2 group, survival and viral RNA load were not statistically distinguishable (Fig. 2d, e) and probably represent natural endpoint variation associated with suboptimal therapeutic effect.

In summary, GS-5734 is a potent and selective inhibitor of EBOV in multiple relevant permissive cell types. In healthy NHPs, intravenous administration of GS-5734 resulted in rapid accumulation and persistence of intracellular NTP. In an NHP model of fatal EVD, pronounced antiviral effects, amelioration of EVD signs, and significant survival benefit was achieved despite treatment initiation on day 3, a time when systemic viral RNA was detectable. These results represent the first case of substantive post-exposure protection against EVD by a small-molecule antiviral compound in NHPs. Intravenous GS-5734 is currently being evaluated in multiple-dose studies in healthy human volunteers to assess clinical safety and pharmacokinetics, and help determine whether GS-5734 may provide therapeutic benefit in acute or recrudescent cases of EVD, or in survivors with prolonged virus shedding and/or chronic clinical sequelae. The broad-spectrum antiviral activity of GS-5734 and its amenability to large-scale production warrants further assessment of its therapeutic potential against other human viral pathogens for which no treatment is available.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 October 2015; accepted 29 January 2016.
Published online 2 March 2016.

1. World Health Organization. *Ebola Situation Report - 3 February 2016*. http://apps.who.int/ebola/current-situation/ebola-situation-report-3-february-2016 (2016).
2. Nanyonga, M., Saidu, J., Ramsay, A., Shindo, N. & Bausch, D. G. Sequelae of Ebola virus disease, Kenema District, Sierra Leone. *Clin. Infect. Dis.* **62,** 125–126 (2016).
3. Varkey, J. B. *et al.* Persistence of Ebola virus in ocular fluid during convalescence. *N. Engl. J. Med.* **372,** 2423–2427 (2015).
4. Mate, S. E. *et al.* Molecular evidence of sexual transmission of Ebola virus. *N. Engl. J. Med.* **373,** 2448–2454 (2015).
5. Deen, G. F. *et al.* Ebola RNA persistence in semen of Ebola virus disease survivors — preliminary report. *N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1511410 (2015).
6. Kuhn, J. H. *Filoviruses: A Compendium of 40 Years of Epidemiological, Clinical, and Laboratory Studies* (SpringWien, 2008).
7. Kupferschmidt, K. & Cohen, J. Infectious diseases. Ebola drug trials lurch ahead. *Science* **347,** 701–702 (2015).
8. Smither, S. J. *et al.* Post-exposure efficacy of oral T-705 (Favipiravir) against inhalational Ebola virus infection in a mouse model. *Antiviral Res.* **104,** 153–155 (2014).
9. Oestereich, L. *et al.* Successful treatment of advanced Ebola virus infection with T-705 (favipiravir) in a small animal model. *Antiviral Res.* **105,** 17–21 (2014).
10. McMullan, L. K. *et al.* The lipid moiety of brincidofovir is required for *in vitro* antiviral activity against Ebola virus. *Antiviral Res.* **125,** 71–78 (2016).
11. Sissoko, D. *et al.* Favipiravir in patients with Ebola virus disease: early results of the JIKI trial in Guinea. *Conference of Retroviruses and Opportunistic Infections* abstr. 103-ALB (Seattle, 2015).
12. Chimerix. *Brincidofovir Will Not Be Considered in Further Clinical Trials in Ebola Virus Disease*. http://ir.chimerix.com/releasedetail.cfm?ReleaseID=893927 (2015).
13. Warren, T. K. *et al.* Protection against filovirus diseases by a novel broad-spectrum nucleoside analogue BCX4430. *Nature* **508,** 402–405 (2014).
14. BioCryst Pharmaceuticals. *BioCryst Announces Study Results for BCX4430 in a Non-Human Primate Model of Ebola Virus Infection*. http://investor.shareholder.com/biocryst/releasedetail.cfm?ReleaseID=888802 (2014).
15. Thi, E. P. *et al.* Lipid nanoparticle siRNA treatment of Ebola-virus-Makona-infected nonhuman primates. *Nature* **521,** 362–365 (2015).
16. Qiu, X. *et al.* Reversion of advanced Ebola virus disease in nonhuman primates with ZMapp. *Nature* **514,** 47–53 (2014).
17. Olinger, G. G., Jr *et al.* Delayed treatment of Ebola virus infection with plant-derived monoclonal antibodies provides protection in rhesus macaques. *Proc. Natl Acad. Sci. USA* **109,** 18030–18035 (2012).
18. Tekmira Pharmaceuticals Corporation. *Tekmira Provides Update on TKM-Ebola-Guinea*. http://www.sec.gov/Archives/edgar/data/1447028/000117184315003522/newsrelease.htm (2015).
19. Cho, A. *et al.* Synthesis and antiviral activity of a series of 1'-substituted 4-aza-7,9-dideazaadenosine C-nucleosides. *Bioorg. Med. Chem. Lett.* **22,** 2705–2707 (2012).
20. Murakami, E. *et al.* The mechanism of action of β-D-2'-deoxy-2'-fluoro-2'-*C*-methylcytidine involves a second metabolic pathway leading to β-D-2'-deoxy-2'-fluoro-2'-*C*-methyluridine 5'-triphosphate, a potent inhibitor of the hepatitis C virus RNA-dependent RNA polymerase. *Antimicrob. Agents Chemother.* **52,** 458–464 (2008).
21. Mackman, R. L., Parrish, J. P., Ray, A. S. & Theodore, D. A. Methods and compounds for treating respiratory syncytial virus infections. US Patent 2011045102. (2011).
22. Jácome, R., Becerra, A., Ponce de Leon, S. & Lazcano, A. Structural analysis of monomeric RNA-dependent polymerases: Evolutionary and therapeutic implications. *PLoS ONE* **10,** e0139001 (2015).
23. Bahar, F. G., Ohura, K., Ogihara, T. & Imai, T. Species difference of esterase expression and hydrolase activity in plasma. *J. Pharm. Sci.* **101,** 3979–3988 (2012).
24. Hunt, L. *et al.* Clinical presentation, biochemical, and haematological parameters and their association with outcome in patients with Ebola virus disease: an observational cohort study. *Lancet Infect. Dis.* **15,** 1292–1299 (2015).
25. Martins, K. *et al.* Characterization of clinical and immunological parameters during Ebola virus infection of rhesus macaques. *Viral Immunol.* **28,** 32–41 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements T. Bocan, A. Duplantier, R. Panchal, and C. Kane provided scientific input. B. Norquist assisted with manuscript preparation. C. Cooper provided scientific input with human macrophage cultures for high-content image assessments. S. Tritsch and G. Gomba assisted with GS-5734 dose preparations for efficacy studies. C. Rice provided animal husbandry support services. X. Wei, W. Garner, and L. Zhong provided additional support for statistical analyses. K. Wang, K. Brendza, T. Alfredson, and L. Serafini assisted with analytical methods; S. Bondy and R. Seemayer procured key raw materials; L. Heumann, R. Polniaszeck, E. Rueden, A. Chtchemelinine, K. Brak, and B. Hoang contributed to synthesis; and Y. Zherebina helped with chiral separations. G. Lee supported the RSV antiviral assay, and G. Stepan, S. Ahmadyar, and H. Yu conducted part of the cytotoxicity testing. J. Knox contributed to polymerase modelling. A. L. Rheingold performed the X-ray crystallographic analysis (Supplementary Information). Studies at USAMRIID were in part supported by The Joint Science and Technology Office for Chemical and Biological Defense (JSTO-CBD) of the Defense Threat Reduction Agency (DTRA) under plan #CB10218. Work in the Fearns laboratory was supported by NIH R01AI113321. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Army or the Centers for Disease Control and Prevention, US Department of Health and Human Services.

Author Contributions R.L.M., D.S., H.C.H., E.D., S.N., E.C., M.O.C., L.Z., W.L., B.S., Q.W., K.C., and L.W. were responsible for the synthesis, characterization, and scale-up of small molecules. T.K.W. designed and supervised activities associated with efficacy evaluations, and interpreted study results. J.W., K.S.S., N.L.G., G.D., S.A.V.T., and J.E.N. conducted *in vivo* efficacy studies and performed associated sample analyses. A.C.S., L.S.W., and L.G. coordinated efficacy study activities. M.K.L., M.F., L.K.M., designed and executed the initial *in vitro* antiviral testing against EBOV and analysed data. V.S., R.Z., C.J.R., D.G, T.K., and B.P.E. designed and executed cell-based infection assays and analysed these data. E.G. conducted quantitative PCR analysis. D.K.N. and C.L.W. performed anatomic pathology examinations and analyses of all nonhuman primate subjects. E.R.N., J.R.K., and G.P. conducted viral genomic sequence analyses. N.L., I.T., and R.S. developed and tested drug formulations. A.S.R., D.B., Y.P., and K.M.S. designed and executed the pharmacokinetic and metabolism studies and summarized results. M.P., O.B., M.R.B., and R.F. designed and conducted biochemical enzymatic assays. K.M.S., J.Y.F., and Y.X. conducted cell-based assays for cytotoxicity. P.W. conducted statistical analysis and S.-S.C. oversaw the analysis. A.S.R., R.J., R.L.M., V.S., R.B., S.S., D.L.M., C.F.S., S.T.N., W.A.L., T.C., and S.B. designed experiments, evaluated results, and provided project oversight. T.K.W., A.S.R., R.J., D.S., M.P., and T.C. outlined and wrote the manuscript.

Author Information Viral genomic sequences have been deposited in GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and accession numbers are supplied in Extended Data Table 5. Small molecule X-ray crystallographic coordinates and structure factor files have been deposited in the Cambridge Structural Database (http://www.ccdc.cam.ac.uk/) and accession numbers are supplied in the Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.B. (sina.bavari.civ@mail.mil) and T.C. (tomas.cihlar@gilead.com).

METHODS

Data reporting. No statistical methods were used to predetermine sample size for biochemical or cell-based assays, or for pharmacokinetic studies. Investigators were not blinded to outcome assessment during these investigations. For GS-5734 efficacy assessments in nonhuman primates, statistical power analysis was used to predetermine sample size, and subjects were randomly assigned to experimental group, stratified by sex and balanced by body weight. Study personnel responsible for assessing animal health (including euthanasia assessment) and administering treatments were experimentally blinded to group assignment of animals and outcome.

Small molecules. GS-5734, Nuc, and NTP were synthesized at Gilead Sciences, Inc., and chemical identity and sample purity were established using NMR, HRMS, and HPLC analysis (Supplementary Information). The radiolabelled analogue [^{14}C]GS-5734 (specific activity, 58.0 mCi mmol^{-1}) was obtained from Moravek Biochemicals (Brea, California) and was prepared in a similar manner described for GS-5734 using [^{14}C]trimethylsilylcyanide (Supplementary Information). Small molecule X-ray crystallographic coordinates and structure factor files have been deposited in the Cambridge Structural Database (http://www.ccdc.cam.ac.uk/) and accession numbers are supplied in the Supplementary Information.

Viruses. RSV A2 was purchased from Advanced Biotechnologies, Inc. EBOV (Kikwit and Makona variants), Sudan virus (SUDV, Gulu), Marburg virus (MARV, Ci67), Junín virus (JUNV, Romero), Lassa virus (LASV, Josiah), Middle East respiratory syndrome virus (MERS, Jordan N3), Chikungunya virus (CHIV, AF 15561), and Venezuelan equine encephalitis virus (VEEV, SH3) were all prepared and characterized at the United States Army Medical Research Institute for infectious diseases (USAMRIID). EBOV containing a GFP reporter gene (EBOV–GFP), EBOV Makona (Liberia, 2014), and MARV containing a GFP reporter gene (MARV–GFP) were prepared and characterized at the Centers for Disease Control and Prevention[26,27].

Cells. HEp-2 (CCL-23), PC-3 (CCL-1435), HeLa (CCL-2), U2OS (HTB-96), Vero (CCL-81), HFF-1 (SCRC-1041), and HepG2 (HB-8065) cell lines were purchased from the American Type Culture Collection. Cell lines were not authenticated and were not tested for mycoplasma as part of routine use in assays. HEp-2 cells were cultured in Eagle's Minimum Essential Media (MEM) with GlutaMAX supplemented with 10% fetal bovine serum (FBS) and 100 U ml^{-1} penicillin and streptomycin. PC-3 cells were cultured in Kaighn's F12 media supplemented with 10% FBS and 100 U ml^{-1} penicillin and streptomycin. HeLa, U2OS, and Vero cells were cultured in MEM supplemented with 10% FBS, 1% L-glutamine, 10 mM HEPES, 1% non-essential amino acids, and 1% penicillin/streptomycin. HFF-1 cells were cultured in MEM supplemented with 10% FBS and 0.5 mM sodium pyruvate. HepG2 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) with GlutaMAX supplemented with 10% FBS, 100 U ml^{-1} penicillin and streptomycin, and 0.1 mM non-essential amino acids. The MT-4 cell line was obtained from the NIH AIDS Research and Reference Reagent Program and cultured in RPMI-1640 medium supplemented with 10% FBS, 100 U ml^{-1} penicillin and streptomycin, and 2 mM L-glutamine. The Huh-7 cell line was obtained from C. M. Rice (Rockefeller University) and cultured in DMEM supplemented with 10% FBS, 100 U ml^{-1} penicillin and streptomycin, and non-essential amino acids.

Primary human hepatocytes were purchased from Invitrogen and cultured in William's Medium E medium containing cell maintenance supplement. Donor profiles were limited to 18- to 65-year-old nonsmokers with limited alcohol consumption. Upon delivery, the cells were allowed to recover for 24 h in complete medium with supplement provided by the vendor at 37 °C. Human PBMCs were isolated from human buffy coats obtained from healthy volunteers (Stanford Medical School Blood Center, Palo Alto, California) and maintained in RPMI-1640 with GlutaMAX supplemented with 10% FBS, 100 U ml^{-1} penicillin and streptomycin. Rhesus fresh whole blood was obtained from Valley Biosystems. PBMCs were isolated from whole blood by Ficoll-Hypaque density gradient centrifugation. Briefly, blood was overlaid on 15 ml Ficoll-Paque (GE Healthcare Bio-Sciences AB), and centrifuged at 500*g* for 20 min. The top layer containing platelets and plasma was removed, and the middle layer containing PBMCs was transferred to a fresh tube, diluted with Tris buffered saline up to 50 ml, and centrifuged at 500*g* for 5 min. The supernatant was removed and the cell pellet was resuspended in 5 ml red blood cell lysis buffer (155 mM ammonium chloride, 10 mM potassium bicarbonate, 0.1 mM EDTA, pH 7.5). To generate stimulated PBMCs, freshly isolated quiescent PBMCs were seeded into a T-150 (150 cm^2) tissue culture flask containing fresh medium supplemented with 10 U ml^{-1} of recombinant human interleukin-2 (IL-2) and 1 μg ml^{-1} phytohaemagglutinin-P at a density of 2×10^6 cells ml^{-1} and incubated for 72 h at 37 °C. Human macrophage cultures were isolated from PBMCs that were purified by Ficoll gradient centrifugation from 50 ml of blood from healthy human volunteers. PBMCs were cultured for 7 to 8 days in in RPMI cell culture media supplemented with 10% FBS, 5 to 50 ng ml^{-1} granulocyte-macrophage colony-stimulating factor and 50 μM β-mercaptoethanol to induce macrophage differentiation. The cryopreserved human primary renal proximal tubule epithelial cells were obtained from LifeLine Cell Technology and isolated from the tissue of human kidney. The cells were cultured at 90% confluency with RenaLife complete medium in a T-75 flask for 3 to 4 days before seeding into 96-well assay plates. Immortalized human microvascular endothelial cells (HMVEC-TERT) were obtained from R. Shao at the Pioneer Valley Life Sciences Institute[28]. HMVEC-TERT cells were cultured in endothelial basal media supplemented with 10% FBS, 5 μg of epithelial growth factor, 0.5 mg hydrocortisone, and gentamycin/amphotericin-B.

Enzymes. RNA POLII was purchased as part of the HeLaScribe Nuclear Extract *in vitro* Transcription System kit from Promega. The recombinant human POLRMT and transcription factors mitochondrial transcription factors A (mtTFA or TFAM) and B2 (mtTFB2 or TFB2M) were purchased from Enzymax. RSV ribonucleoprotein (RNP) complexes were prepared according to a method modified from ref. 29.

Intracellular metabolism studies. The intracellular metabolism of GS-5734 was assessed in different cell types (HMVEC and HeLa cell lines, and primary human and rhesus PBMCs, monocytes and monocyte-derived macrophages) following 2-h pulse or 72-h continuous incubations with 10 μM GS-5734. For comparison, intracellular metabolism during a 72-h incubation with 10 μM of Nuc was completed in human monocyte-derived macrophages. For pulse incubations, monocyte-derived macrophages isolated from rhesus monkeys or humans were incubated for 2 h in compound-containing media followed by removal, washing with 37 °C drug-free media, and incubated for an additional 22 h in media which did not contain GS-5734. Human monocyte-derived macrophages, HeLa and HMVEC were grown to confluence (approximately 0.5, 0.2, and 1.2×10^6 cells per well, respectively) in 500 μl of media in 12-well tissue culture plates. Monocyte and PBMCs were incubated in suspension (approximately 1×10^6 cells ml^{-1}) in 1 ml of media in micro centrifuge tubes.

For adherent cells (HMVEC, HeLa, and monocyte-derived macrophages), media was removed at select time points from duplicate wells, cells washed twice with 2 ml of ice-cold 0.9% normal saline. For non-adherent cells (monocytes and PBMCs), duplicate incubations were centrifuged at 2,500*g* for 30 s to remove media. The cell pellets were re-suspended with 500 μl cell culture media (RPMI with 10% FBS) and layered on top of a 500 μl oil layer (Nyosil M25; Nye Lubricants) in a microcentrifuge tube. Samples were then centrifuged at room temperature at 13,000 r.p.m. for 45 s. The media layer was removed and the oil layer was washed twice with 500 μl water. The oil layer was then carefully removed using a Pasteur pipet attached to vacuum. A volume of 0.5 ml of 70% methanol containing 100 nM of the analytical internal standard 2-chloro-adenosine-5′-triphosphate (Sigma-Aldrich) was added to isolated cells. Samples were stored overnight at −20 °C to facilitate extraction, centrifuged at 15,000*g* for 15 min and then supernatant was transferred to clean tubes for drying in a MiVac Duo concentrator (Genevac). Dried samples were then reconstituted in mobile phase A containing 3 mM ammonium formate (pH 5.0) with 10 mM dimethylhexylamine (DMH) in water for analysis by liquid chromatography coupled to triple quadrupole mass spectrometry (LC-MS/MS).

LC-MS/MS was performed using low-flow ion-pairing chromatography, similar to methods described previously[30]. Briefly, analytes were separated using a 50 × 2 mm × 2.5 μm Luna C18(2) HST column (Phenomenex) connected to a LC-20ADXR (Shimadzu) ternary pump system and HTS PAL autosampler (LEAP Technologies). A multi-stage linear gradient from 10% to 50% acetonitrile in a mobile phase containing 3 mM ammonium formate (pH 5.0) with 10 mM dimethylhexylamine over 8 min at a flow rate of 150 μl min^{-1} was used to separate analytes. Detection was performed on an API 4000 (Applied Biosystems) MS/MS operating in positive ion and multiple reaction monitoring modes. Intracellular metabolites alanine metabolite, Nuc, nucleoside monophosphate, nucleoside diphosphate, and nucleoside triphosphate were quantified using 7-point standard curves ranging from 0.274 to 200 pmol (approximately 0.5 to 400 μM) prepared in cell extract from untreated cells. Levels of adenosine nucleotides were also quantified to assure dephosphorylation had not taken place during sample collection and preparation. In order to calculate intracellular concentration of metabolites, the total number of cells per sample were counted using a Countess automated cell counter (Invitrogen).

EBOV Huh-7 and HMVEC antiviral assay. Antiviral assays were conducted in biosafety level 4 containment (BSL-4) at the Centers for Disease Control and Prevention. EBOV antiviral assays were conducted in primary HMVEC-TERT and in Huh-7 cells. Huh-7 cells were not authenticated and were not tested for mycoplasma. Ten concentrations of compound were diluted in fourfold serial dilution increments in media, and 100 μl per well of each dilution was transferred in duplicate (Huh-7) or quadruplicate (HMVEC-TERT) onto 96-well assay plates containing cell monolayers. The plates were transferred to BSL-4 containment, and the appropriate dilution of virus stock was added to test plates containing cells and serially diluted compounds. Each plate included four wells of infected

untreated cells and four wells of uninfected cells that served as 0% and 100% virus inhibition controls, respectively. After the infection, assay plates were incubated for 3 days (Huh-7) or 5 days (HMVEC-TERT) in a tissue culture incubator. Virus replication was measured by direct fluorescence using a Biotek HTSynergy plate reader. For virus yield assays, Huh-7 cells were infected with wild-type EBOV for 1 h at 0.1 plaque-forming units (PFU) per cell. The virus inoculum was removed and replaced with 100 μl per well of media containing the appropriate dilution of compound. At 3 days post-infection, supernatants were collected, and the amount of virus was quantified by endpoint dilution assay. The endpoint dilution assay was conducted by preparing serial dilutions of the assay media and adding these dilutions to fresh Vero cell monolayers in 96-well plates to determine the tissue culture infectious dose that caused 50% cytopathic effects ($TCID_{50}$). To measure levels of viral RNA from infected cells, total RNA was extracted using the MagMAX-96 Total RNA Isolation Kit and quantified using a quantitative reverse transcription polymerase chain reaction (qRT–PCR) assay with primers and probes specific for the EBOV nucleoprotein gene.

EBOV assay in HeLa and HFF-1 cells. Antiviral assays were conducted in BSL-4 at USAMRIID. HeLa or HFF-1 cells were seeded at 2,000 cells per well in 384-well plates. Ten serial dilutions of compound in triplicate were added directly to the cell cultures using the HP D300 digital dispenser (Hewlett Packard) in twofold dilution increments starting at 10 μM at 2 h before infection. The DMSO concentration in each well was normalized to 1% using an HP D300 digital dispenser. The assay plates were transferred to the BSL-4 suite and infected with EBOV Kikwit at a multiplicity of infection of 0.5 PFU per cell for HeLa cells and with EBOV Makona at a multiplicity of infection of 5 PFU per cell for HFF-1 cells. The assay plates were incubated in a tissue culture incubator for 48 h. Infection was terminated by fixing the samples in 10% formalin solution for an additional 48 h before immune-staining, as described in Supplementary Table 1.

EBOV human macrophage infection assay. Antiviral assays were conducted in BSL-4 at USAMRIID. Primary human macrophage cells were seeded in a 96-well plate at 40,000 cells per well. Eight to ten serial dilutions of compound in triplicate were added directly to the cell cultures using an HP D300 digital dispenser in threefold dilution increments 2 h before infection. The concentration of DMSO was normalized to 1% in all wells. The plates were transferred into the BSL-4 suite, and the cells were infected with 1 PFU per cell of EBOV in 100 μl of media and incubated for 1 h. The inoculum was removed, and the media was replaced with fresh media containing diluted compounds. At 48 h post-infection, virus replication was quantified by immuno-staining as described in Supplementary Table 1.

RSV A2 antiviral assay. For antiviral tests, compounds were threefold serially diluted in source plates from which 100 nl of diluted compound was transferred to a 384-well cell culture plate using an Echo acoustic transfer apparatus. HEp-2 cells were added at a density of 5×10^5 cells per ml, then infected by adding RSV A2 at a titer of $1 \times 10^{4.5}$ tissue culture infectious doses ($TCID_{50}$) per ml. Immediately following virus addition, 20 μl of the virus and cells mixture was added to the 384-well cell culture plates using a μFlow liquid dispenser and cultured for 4 days at 37 °C. After incubation, the cells were allowed to equilibrate to 25 °C for 30 min. The RSV-induced cytopathic effect was determined by adding 20 μl of CellTiter-Glo Viability Reagent. After a 10-min incubation at 25 °C, cell viability was determined by measuring luminescence using an Envision plate reader.

High content imaging assay detecting viral infection. Antiviral assays were conducted in 384-or 96-well plates in BSL-4 at USAMRIID using a high-content imaging system to quantify virus antigen production as a measure of virus infection. A 'no virus' control and a '1% DMSO' control were included to determine the 0% and 100% virus infection, respectively. The primary and secondary antibodies and dyes used for nuclear and cytoplasmic staining are listed in Supplementary Table 1. The primary antibody specific for a particular viral protein was diluted 1,000-fold in blocking buffer (1 × PBS with 3% BSA) and added to each well of the assay plate. The assay plates were incubated for 60 min at room temperature. The primary antibody was removed, and the cells were washed three times with 1 × PBS. The secondary detection antibody was an anti-mouse (or rabbit) IgG conjugated with Dylight488 (Thermo Fisher Scientific, catalogue number 405310). The secondary antibody was diluted 1,000-fold in blocking buffer and was added to each well in the assay plate. Assay plates were incubated for 60 min at room temperature. Nuclei were stained using Draq5 (Biostatus) or 33342 Hoechst (ThermoFisher Scientific) for Vero and HFF-1 cell lines. Both dyes were diluted in 1× PBS. The cytoplasm of HFF-1 (EBOV assay) and Vero E6 (MERS assay) cells were counter-stained with CellMask Deep Red (Thermo Fisher Scientific). Cell images were acquired using a Perkin Elmer Opera confocal plate reader (Perkin Elmer) using a ×10 air objective to collect five images per well. Virus-specific antigen was quantified by measuring fluorescence emission at a 488 nm wavelength and the stained nuclei were quantified by measuring fluorescence emission at a 640 nm wavelength. Acquired images were analysed using Harmony and Acapella PE software. The Draq5 signal was used to generate a nuclei mask to define each nuclei in the image for quantification of cell number. The CellMask Deep Red dye was used to demarcate the Vero and HFF-1 cell borders for cell-number quantitation. The viral-antigen signal was compartmentalized within the cell mask. Cells that exhibited antigen signal higher than the selected threshold were counted as positive for viral infection. The ratio of virus-positive cells to total number of analysed cells was used to determine the percentage of infection for each well on the assay plates. The effect of compounds on the viral infection was assessed as percentage of inhibition of infection in comparison to control wells. The resultant cell number and percentage of infection were normalized for each assay plate. Analysis of dose–response curve was performed using GeneData Screener software applying Levenberg–Marquardt algorithm for curve-fitting strategy. The curve-fitting process, including individual data point exclusion, was pre-specified by default software settings. R^2 value quantified goodness of fit and fitting strategy was considered acceptable at $R^2 > 0.8$.

Virus assays. All virus infections were quantified by immuno-staining using antibodies that recognized the relevant viral glycoproteins, as described in Supplementary Table 1.

Marburg virus assay. HeLa cells were seeded at 2,000 cells per well in a 384-well plate, and compounds were added to the assay plates. Assay plates were transferred to the BSL-4 suite and infected with 1 PFU per cell MARV, which resulted in 50% to 70% of the cells expressing virus antigen in a 48-h period.

Sudan virus assay. HeLa cells were seeded at 2,000 cells per well in a 384-well plate, and compounds were added to the assay plates. Assay plates were transferred to the BSL-4 suite and infected with 0.08 PFU SUDV per cell, which resulted in 50% to 70% of the cells expressing virus antigen in a 48-h period.

Junín virus assay. HeLa cells were seeded at 2,000 cells per well in a 384-well plate, and compounds were added to the assay plates. Assay plates were transferred to the BSL-4 suite and infected with 0.3 PFU per cell JUNV, which resulted in ~50% of the cells expressing virus antigen in a 48-h period.

Lassa fever virus assay. HeLa cells were seeded at 2,000 cells per well in a 384-well plate, and compounds were added to the assay plates. Assay plates were transferred to the BSL-4 suite and infected with 0.1 PFU per cell LASV, which resulted in >60% of the cells expressing virus antigen in a 48-h period.

Middle East respiratory syndrome virus assay. African green monkey (*Chlorocebus* sp.) kidney epithelial cells (Vero E6) were seeded at 4,000 cells per well in a 384-well plate, and compounds were added to the assay plates. Assay plates were transferred to the BSL-4 suite and infected with 0.5 PFU per cell of MERS virus, which resulted in >70% of the cells expressing virus antigen in a 48-h period.

Chikungunya virus assay. U2OS cells were seeded at 3,000 cells per well in a 384-well plate, and compounds were added to the assay plates. Assay plates were transferred to the BSL-4 suite and infected with 0.5 PFU per cell of CHIK, which resulted in >80% of the cells expressing virus antigen in a 48-h period.

Venezuelan equine encephalitis virus assay. HeLa cells were seeded at 4,000 cells per well in a 384-well plate, and compounds were added to the assay plates. Assay plates were transferred to the BSL-4 suite and infected with 0.1 PFU per cell VEEV, which resulted in >60% of the cells expressing virus antigen in a 20-h period.

Cytotoxicity assays. HEp-2 (1.5×10^3 cells per well) and MT-4 (2×10^3 cells per well) cells were plated in 384-well plates and incubated with the appropriate medium containing threefold serially diluted compound ranging from 15 nM to 100,000 nM. PC-3 cells (2.5×10^3 cells per well), HepG2 cells (4×10^3 cells per well), hepatocytes (1×10^6 cells per well), quiescent PBMCs (1×10^6 cells per well), stimulated PBMCs (2×10^5 cells per well), and RPTEC cells (1×10^3 cells per well) were plated in 96-well plates and incubated with the appropriate medium containing threefold serially diluted compound ranging from 15 nM to 100,000 nM. Cells were cultured for 4–5 days at 37 °C. Following the incubation, the cells were allowed to equilibrate to 25 °C, and cell viability was determined by adding Cell-Titer Glo viability reagent. The mixture was incubated for 10 min, and the luminescence signal was quantified using an Envision plate reader. Cell lines were not authenticated and were not tested for mycoplasma as part of routine use in cytotoxicity assays.

***In vitro* RSV RNA synthesis assay.** RNA synthesis by the RSV polymerase was reconstituted *in vitro* using purified RSV L/P complexes and an RNA oligonucleotide template (Dharmacon), representing nucleotides 1–14 of the RSV leader promoter[31–33] (3′-UGCGCUUUUUUACG-5′). RNA synthesis reactions were performed as described previously, except that the reaction mixture contained 250 μM guanosine triphosphate (GTP), 10 μM uridine triphosphate (UTP), 10 μM cytidine triphosphate (CTP), supplemented with 10 μCi [α-^{32}P]CTP, and either included 10 μM adenosine triphosphate (ATP) or no ATP. Under these conditions, the polymerase is able to initiate synthesis from the position 3 site of the promoter, but not the position 1 site. The NTP metabolite of GS-5734 was serially diluted in DMSO and included in each reaction mixture at concentrations of 10, 30, or 100 μM as specified in Fig. 1f. RNA products were analysed by electrophoresis on a 25% polyacrylamide gel, containing 7 M urea, in Tris–taurine–EDTA buffer, and radiolabelled RNA products were detected by autoradiography.

RSV A2 polymerase inhibition assay. Transcription reactions contained 25 μg of crude RSV RNP complexes in 30 μL of reaction buffer (50 mM Tris-acetate (pH 8.0), 120 mM potassium acetate, 5% glycerol, 4.5 mM $MgCl_2$, 3 mM DTT, 2 mM EGTA, 50 μg ml^{-1} BSA, 2.5 U RNasin, 20 μM ATP, 100 μM GTP, 100 μM UTP, 100 μM CTP, and 1.5 μCi [α-^{32}P]ATP (3,000 Ci $mmol^{-1}$)). The radiolabelled nucleotide used in the transcription assay was selected to match the nucleotide analogue being evaluated for inhibition of RSV RNP transcription.

To determine whether nucleotide analogues inhibited RSV RNP transcription, compounds were added using a six-step serial dilution in fivefold increments. After a 90-min incubation at 30 °C, the RNP reactions were stopped with 350 μl of Qiagen RLT lysis buffer, and the RNA was purified using a Qiagen RNeasy 96 kit. Purified RNA was denatured in RNA sample loading buffer at 65 °C for 10 min and run on a 1.2% agarose/MOPS gel containing 2 M formaldehyde. The agarose gel was dried, exposed to a Storm phosphorimaging screen, and developed using a Storm phosphorimager.

Inhibition of human RNA polymerase II. For a 25 μl reaction mixture, 7.5 μl 1 × transcription buffer (20 mM HEPES (pH 7.2–7.5), 100 mM KCl, 0.2 mM EDTA, 0.5 mM DTT, 20% glycerol), 3 mM $MgCl_2$, 100 ng CMV positive or negative control DNA, and a mixture of ATP, GTP, CTP and UTP was pre-incubated with various concentrations (0–500 μM) of the inhibitor at 30 °C for 5 min. The mixture contained 5–25 μM (equal to K_m) of the competing ^{33}P-labelled ATP and 400 μM of GTP, UTP, and CTP. The reaction was started by addition of 3.5 μl of HeLa and extract. After 1 h of incubation at 30 °C, the polymerase reaction was stopped by addition of 10.6 μl proteinase K mixture that contained final concentrations of 2.5 μg $μl^{-1}$ proteinase K, 5% SDS, and 25 mM EDTA. After incubation at 37 °C for 3–12 h, 10 μl of the reaction mixture was mixed with 10 μl of the loading dye (98% formamide, 0.1% xylene cyanol and 0.1% bromophenol blue), heated at 75 °C for 5 min, and loaded onto a 6% polyacrylamide gel (8 M urea). The gel was dried for 45 min at 70 °C and exposed to a phosphorimager screen. The full length product, 363 nucleotide runoff RNA, was quantified using a Typhoon Trio Imager and Image Quant TL Software.

Inhibition of human mitochondrial RNA polymerase. Twenty nanomolar POLRMT was incubated with 20 nM template plasmid (pUC18-LSP) containing POLRMT light-strand promoter region and mitochondrial (mt) transcription factors TFA (100 nM) and mtTFB2 (20 nM) in buffer containing 10 mM HEPES (pH 7.5), 20 mM NaCl, 10 mM DTT, 0.1 mg ml^{-1} BSA, and 10 mM $MgCl_2$[34]. The reaction mixture was pre-incubated to 32 °C, and the reactions were initiated by addition of 2.5 μM of each of the natural NTPs and 1.5 μCi of [^{32}P]GTP. After incubation for 30 min at 32 °C, reactions were spotted on DE81 paper and quantified.

Molecular modelling. A homology model of RSV A2 and EBOV polymerases were built using the HIV reverse transcriptase X-ray crystal structure (PDB:1RTD). Schrödinger Release 2015-1: Prime, version 3.9 (Schrödinger, LLC), default settings with subsequent rigid body minimization and side-chain optimization. Loop insertions not in 1RTD of greater than 10 amino acids were not built.

qRT– PCR for *in vivo* studies. For quantitative assessment of viral RNA non-human primate plasma samples, whole blood was collected using a K3 EDTA Greiner Vacuette tube (or equivalent) and sample centrifuged at 2500 (± 200) relative centrifugal force for 10 ± 2 min. To inactivate virus, plasma was treated with 3 parts (300 μl) TriReagent LS and samples were transferred to frozen storage (−60 °C to −90 °C), until removal for RNA extraction. Carrier RNA and QuantiFast High Concentration Internal Control (Qiagen) were spiked into the sample before extraction, conducted according to manufacturer's instructions. The viral RNA was eluted in AVE buffer. Each extracted RNA sample was tested with the QuantiFast Internal Control RT–PCR RNA Assay (Qiagen) to evaluate the yield of the spiked-in QuantiFast High Concentration Internal Control. If the internal control amplified within manufacturer-designated ranges, further quantitative analysis of the viral target was performed. RT–PCR was conducted using an ABI 7500 Fast Dx using primers specific to EBOV glycoprotein. Samples were run in triplicate using a 5 μl template volume. For quantitative assessments, the average of the triplicate genomic equivalents (GE) per reaction were determined and multiplied by 800 to obtain GE ml^{-1} plasma. Standard curves were generated using synthetic RNA. The limits of quantification for this assay are $8.0 \times 10^4 - 8.0 \times 10^{10}$ GE ml^{-1} of plasma. Acceptance criteria for positive template control (PTC), negative template control (NTC), negative extraction control (NEC), and positive extraction control (PEC) are specified by standard operating procedure. For qualitative assessments, the limit of detection (LOD) was defined as C_t 38.07, based on method validation testing. An animal was considered to have tested positive for detection of EBOV RNA when a minimum of 2 of 3 replicates were designated as 'positive' and PTC, NTC, and NEC controls met specified method-acceptance criteria. A sample was designated as 'positive' when the C_t value was <LOD C_t.

Pharmacokinetic evaluations. Three uninfected male rhesus monkeys (*Macaca mulatta*) were used for the pharmacokinetic study. GS-5734 was formulated in solution at 5 mg ml^{-1} with 12% sulfobutylether-β-cyclodextrin in water, pH 3.5–4.0, and 2 ml kg^{-1} was administered by slow bolus (approximately 1 min) for a final dose of 10 mg kg^{-1}. Blood samples for plasma and PBMCs were collected from a femoral vein/artery and were taken from each monkey over a 24-h period. Plasma samples were obtained at predose and at 0.083, 0.25, 0.5, 1, 2, 4, 8, and 24 h postdose. PBMC samples were obtained at 2, 4, 8, and 24 h. Blood samples for plasma were collected into chilled collection tubes containing sodium fluoride/potassium oxalate as the anticoagulant and were immediately placed on wet ice, followed by centrifugation to obtain plasma. Blood samples for PBMC isolation were collected at room temperature into CPT vacutainer tubes containing sodium heparin for isolation. Plasma and isolated PBMC samples were frozen immediately and stored at ≤60 °C until analysed.

For plasma analysis, an aliquot of 25 μl of each plasma sample was treated with 100 μl of 90% methanol and acetonitrile mixture (1:1, v:v) and 10% water with 20 nM 5-(2-aminopropyl)indole as an internal standard. Then, 100 μl of samples were filtered through an Agilent Captiva 96 well 0.2 μm filter plate. Filtered samples were dried down completely for approximately 20 min and reconstituted with 1% acetonitrile and 99% water with 0.01% formic acid. An aliquot of 10 μl was injected for LC-MS/MS using a HTC Pal autosampler. Analytes were separated on a Phenomenex Synergi Hydro-RP 30A column (75 × 2.0 mM, 4.0 μm) using a Waters Acquity ultra performance LC (Waters Corporation, Milford, MA, USA), a flow rate of 0.26 ml min^{-1}, and a gradient from Mobile phase A containing 0.2% formic acid in 99% water and 1% acetonitrile to mobile phase B containing 0.2% formic acid in 95% acetonitrile and 5% water over 4.5 min. For MS/MS analysis, we used a Waters Xevo TQ-S in positive multiple reaction monitoring mode using an electrospray probe. Plasma concentrations of GS-734, alanine metabolite and Nuc were determined using an 8-point calibration curve spanning a concentration range of over three orders of magnitude. Quality control samples were run at the beginning and end of the run to ensure accuracy and precision within 20%. Intracellular metabolites in PBMCs were quantified by LC-MS/MS as described above for *in vitro* activation studies.

Radiolabelled tissue distribution. Six cynomolgus monkeys (*Macaca fascicularis*) were administered a single dose of [^{14}C]GS-5734 at 10 mg kg^{-1} (25 μCi kg^{-1}) by intravenous administration (slow bolus). Tissues were collected from three animals at 4 and 168 h postdose. The tissues were excised, rinsed with saline, blotted dry, weighed, and placed on wet ice. Tissues (testes, epididymis, eyes and brain; following homogenization) and plasma were analysed by liquid scintillation counting. Concentrations were converted to ng equivalents of GS-5734 per gram of sample.

***In vivo* efficacy.** Rhesus monkeys (*Macaca mulatta*) were challenged on day 0 by intramuscular injection with a target dose of 1,000 PFU of EBOV Kikwit (Ebola virus H. sapiens-tc/COD/1995/Kikwit), which was derived from a clinical specimen obtained during an outbreak occurring in the Democratic Republic of the Congo (formerly Zaire) in 1995. Challenge virus was propagated from the clinical specimen using cultured cells (Vero or Vero E6) for a total of four passages. Animals (3–6 years old) were randomly assigned to experimental treatment groups, stratified by sex (with equal number of males and females per group) and balanced by body weight, using SAS statistical software. Study personnel responsible for assessing animal health (including euthanasia assessment) and administering treatments were experimentally blinded to group assignment of animals. The primary endpoint for efficacy studies was survival to day 28 following virus challenge. GS-5734 was formulated at Gilead Sciences in water with 12% sulfobutylether-β-cyclodextrin (SBE- β-CD), pH adjusted to 3.0 using HCl. Formulations were administered to anaesthetized animals by bolus intravenous injection at a rate of approximately 1 min per dose in the right or left saphenous vein. The volume of all vehicle or GS-5734 injections was 2.0 ml kg^{-1} body weight. Animals were anaesthetized using intramuscular injection of a solution containing ketamine (100 mg ml^{-1}) and acepromazine (10 mg ml^{-1}) at 0.1 ml kg^{-1} body weight.

Animals were observed at least twice daily to monitor for disease signs, and animals that survived to day 28 were deemed to be protected. Study personnel alleviated unnecessary suffering of infected animals by euthanizing clinically moribund animals. The criteria used as the basis for euthanasia of moribund animals were defined before study initiation and included magnitude of responsiveness, reduced body temperature, and/or specified alterations to serum chemistry parameters[35]. Serum chemistry was analysed using a Vitros 350 Chemistry System (Ortho Clinical Diagnostics), and coagulation parameters were evaluated using a Sysmex CA-1500 coagulation analyser (Siemens Healthcare Diagnostics). Haematology analysis was conducted using a Siemens Advia 120 Hematology System with multispecies software (Siemens Healthcare Diagnostics). On days in which GS-5734 or vehicle dosing were scheduled with blood sample collection for clinical pathology or viraemia analysis, blood samples were collected immediately before dose administration.

Viral genomic sequence analysis. Analysis of viral genomic sequence was conducted with the purpose of evaluating genomic sequence change patterns consistent with development of resistance against GS-5734. Attempts were made to obtain

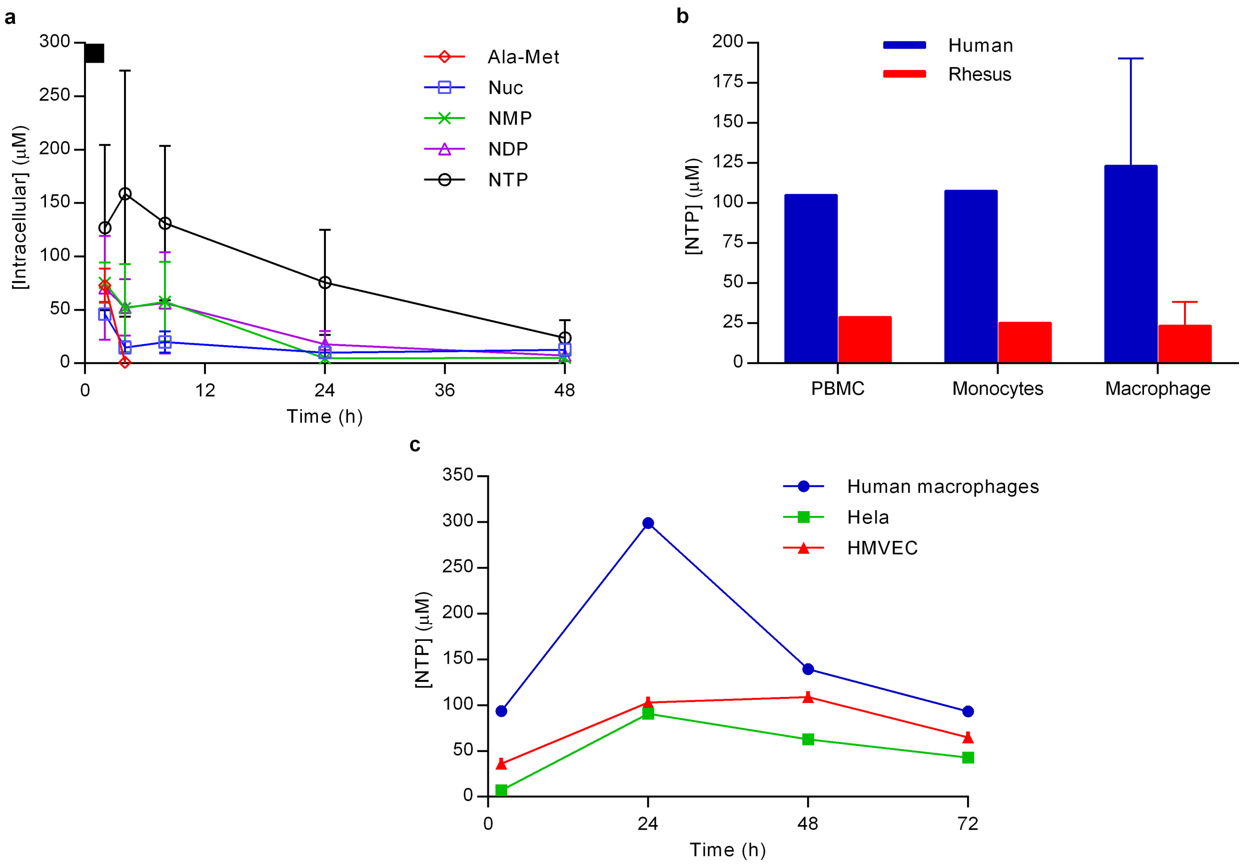
sequence of RNA-dependent RNA polymerase gene (*L*) of Ebola virus from all EBOV-RNA-positive rhesus monkey plasma samples obtained during the efficacy studies. Deep sequencing screening of the *L* gene was completed using previously described methods[36,37]. Mutations and large subclonal events (≥10% of population) were reviewed for: non-synonymous substitutions in treated animals that succumbed to infection, any substitution enriched in treated populations regardless of survival status, and clusters of substitutions in any treated animal.

cDNA synthesis was performed using Superscript III First-Strand Synthesis System (Invitrogen). cDNA was amplified with Phusion Hot Start Flex DNA Polymerase (New England Biolabs) using overlapping 1,500-kb amplicons (primer information available upon request). After pooling and purification with AMPure XP Reagent (Beckman Coulter), PCR products were fragmented using the Covaris S2 instrument (Covaris). Libraries were prepared with the Illumina TruSeq DNA Sample Preparation kit (Illumina) on the Caliper ScicloneG3 Liquid Handling Station (PerkinElmer). After measurement by real-time PCR with the KAPA qPCR Kit (Kapa Biosystems), libraries were diluted to 4 or 10 nM. Cluster amplification was performed on the Illumina cBot, and libraries were sequenced on the Illumina Nextseq or Illumina HiSeq 2500 using the 150 or 100 bp paired-end format. Viral assemblies were completed in DNAStar Lasergene nGen. Amplification primer removal, quality trimming, and trim-to-mer were performed on reads with a minimum similarity to the reference of 93% (four-base mismatch). A target depth of 1,200 is sought and single nucleotide polymorphisms at positions with fewer than 200 read depth were removed from the analysis. A consensus change was defined as a change relative to the deposited EBOV sequence (GenBank: AY354458) present in ≥50% of the population. Below that threshold, single nucleotide polymorphisms were considered subclonal substitutions and part of a minority subpopulation of the virus. Consensus sequence for the region covered in this screening is available in GenBank and the accession numbers are provided in Extended Data Table 5.

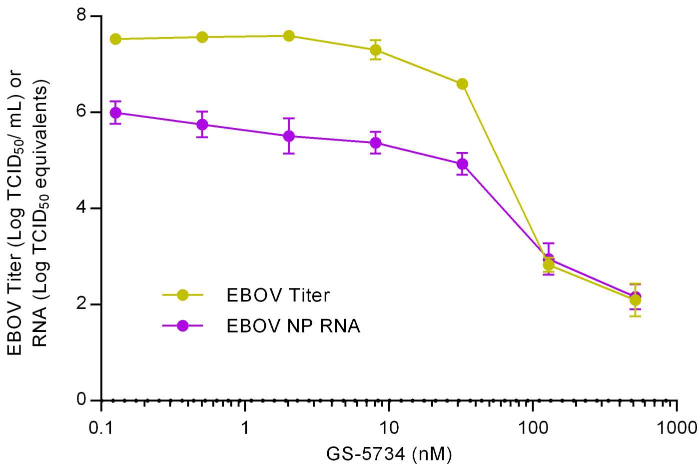
Animal care. Pharmacokinetic and radiolabelled tissue distribution studies in uninfected cynomolgus and rhesus macaques were conducted at Covance, Inc. Protocols were reviewed by an Institutional Animal Care and Use Committee (IACUC) at Covance. Efficacy experiments involving EBOV were performed in animal biosafety level 4 (ABSL-4) at USAMRIID. Research was conducted under an Institutional Animal Care and Use Committee approved protocol in compliance with the Animal Welfare Act, PHS Policy, and other federal statutes and regulations relating to animals and experiments involving animals. The facilities where this research was conducted are accredited by the Association for Assessment and Accreditation of Laboratory Animal Care, International and strictly adhere to principles stated in the Guide for the Care and Use of Laboratory Animals, National Research Council, 2011 (National Academies Press, Washington, DC).

Statistics. Combined vehicle group from part 1 and 2 ($n = 6$ animals total) was used as control group in all statistical comparisons of GS-5734 efficacy evaluations. The impact of GS-5734 treatment on the survival rates was estimated using Kaplan–Meier method and analysed by log-rank analysis using Dunnett–Hsu procedure to adjust for multiple comparisons. The effect on systemic viral RNA levels was assessed by analysis of variance (ANOVA), comparing each GS-5734 treatment group with vehicle group using Dunnett's test to adjust for multiple comparisons. Wilcoxon rank-sum test without adjustment for multiple comparisons was used to compare the effects of GS-5734 treatment on haematology, coagulation, and clinical chemistry parameters. All data met the statistical assumptions of the test performed.

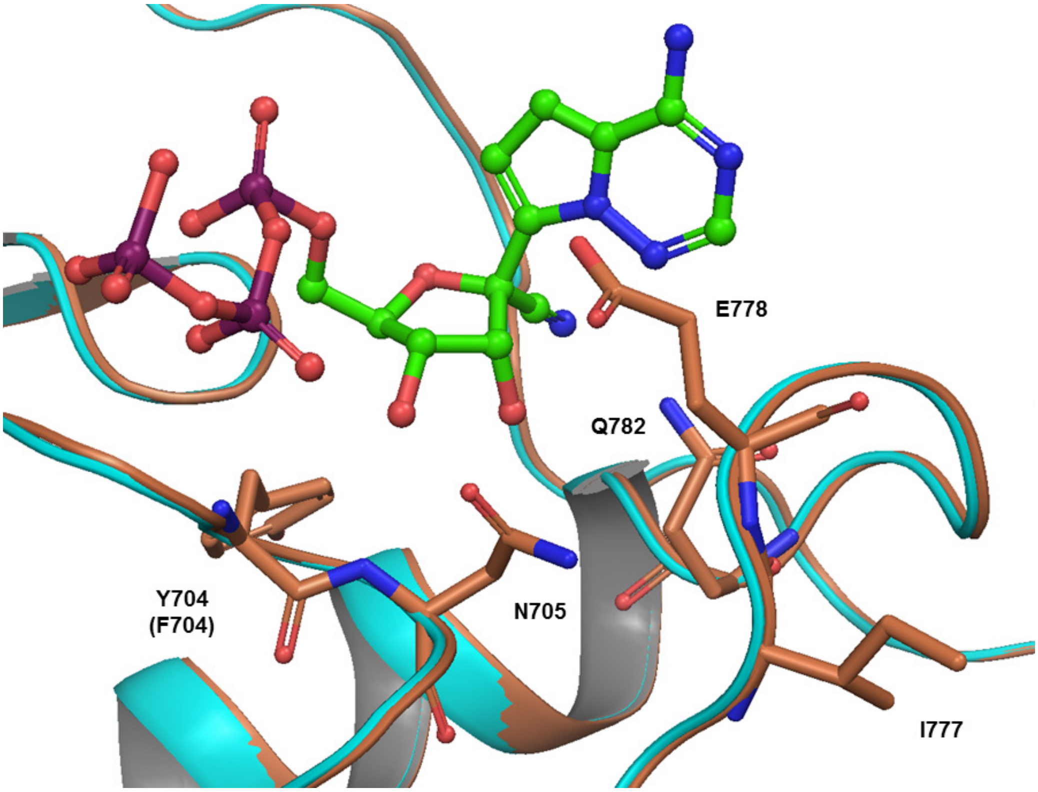
26. Uebelhoer, L. S. *et al.* High-throughput, luciferase-based reverse genetics systems for identifying inhibitors of Marburg and Ebola viruses. *Antiviral Res.* **106,** 86–94 (2014).
27. Towner, J. S. *et al.* Generation of eGFP expressing recombinant Zaire Ebolavirus for analysis of early pathogenesis events and high-throughput antiviral drug screening. *Virology* **332,** 20–27 (2005).
28. Shao, R. & Guo, X. Human microvascular endothelial cells immortalized with human telomerase catalytic protein: a model for the study of in vitro angiogenesis. *Biochem. Biophys. Res. Commun.* **321,** 788–794 (2004).
29. Mason, S. W. *et al.* Polyadenylation-dependent screening assay for respiratory syncytial virus RNA transcriptase activity and identification of an inhibitor. *Nucleic Acids Res.* **32,** 4758–4767 (2004).
30. Durand-Gasselin, L. *et al.* Nucleotide analogue prodrug tenofovir disoproxil enhances lymphoid cell loading following oral administration in monkeys. *Mol. Pharm.* **6,** 1145–1151 (2009).
31. Noton, S. L., Deflube, L. R., Tremaglio, C. Z. & Fearns, R. The respiratory syncytial virus polymerase has multiple RNA synthesis activities at the promoter. *PLoS Pathog.* **8,** e1002980 (2012).
32. Noton, S. L. *et al.* Respiratory syncytial virus inhibitor AZ-27 differentially inhibits different polymerase activities at the promoter. *J. Virol.* **89,** 7786–7798 (2015).
33. Tremaglio, C. Z., Noton, S. L., Deflube, L. R. & Fearns, R. Respiratory syncytial virus polymerase can initiate transcription from position 3 of the leader promoter. *J. Virol.* **87,** 3196–3207 (2013).
34. Lodeiro, M. F. *et al.* Identification of multiple rate-limiting steps during the human mitochondrial transcription cycle in vitro. *J. Biol. Chem.* **285,** 16387–16402 (2010).
35. Warren, T. K. *et al.* Euthanasia assessment in Ebola virus infected nonhuman primates. *Viruses* **6,** 4666–4682 (2014).
36. Kugelman, J. R. *et al.* Emergence of Ebola Virus escape variants in infected nonhuman primates treated with the MB-003 antibody cocktail. *Cell Rep.* **12,** 2111–2120 (2015).
37. Kugelman, J. R. *et al.* Ebola virus genome plasticity as a marker of its passaging history: a comparison of in vitro passaging to non-human primate infection. *PLoS ONE* **7,** e50316 (2012).



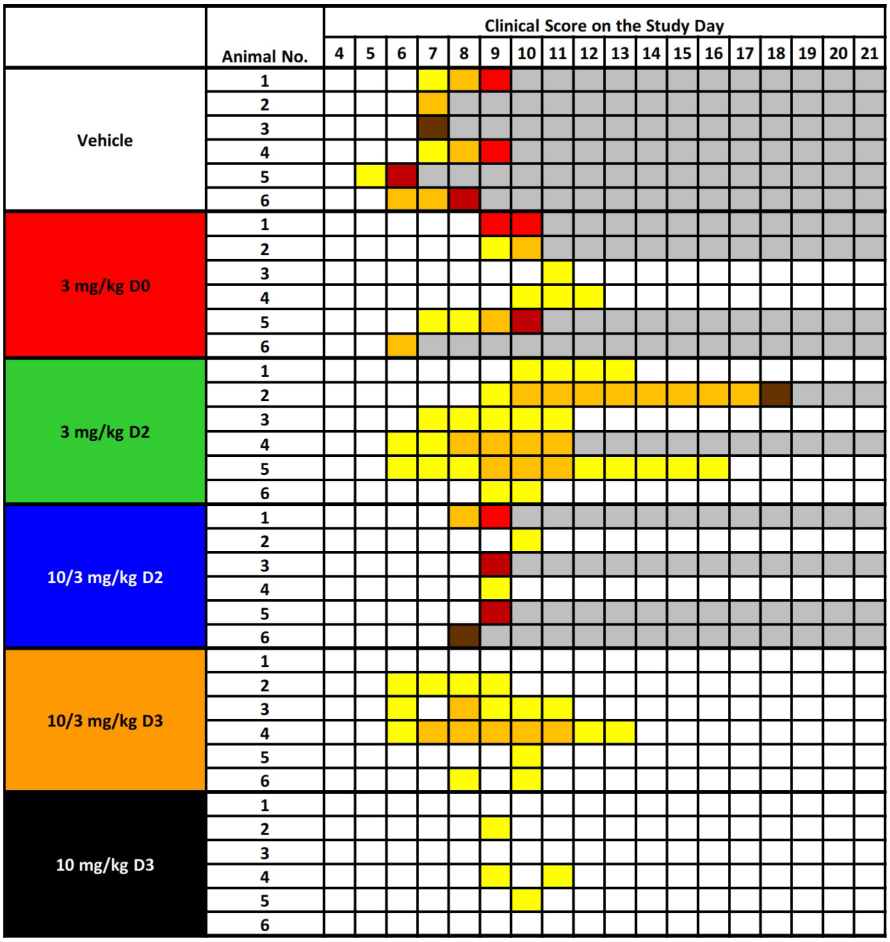
Extended Data Figure 1 | Intracellular metabolism of GS-5734. **a**, Intracellular metabolite profile in human macrophages. Following a 2-h pulse incubation (black bar at top of *y* axis) of human monocyte-derived macrophages with 1 μM GS-5734 (mean ± s.d., from three donors). GS-5734 is rapidly metabolized and not detected in cells. Transient exposure to the intermediate alanine metabolite (Ala-Met) is observed, followed by persistent Nuc exposure. The pharmacologically active NTP is formed quickly, achieving a maximum intracellular concentration at 4 h and persisted with a half-life of 16 ± 1 h in the three donors. Intracellular concentration was estimated on the basis of an intracellular volume of 1 pl per cell. **b**, Efficiency of GS-5734 activation in human and rhesus cells *in vitro*. Intracellular NTP concentrations formed in human and rhesus PBMCs, monocytes, and monocyte-derived macrophages during a 2-h incubation with 1 μM GS-5734 (results are the mean ± s.d. of two (PBMC and monocyte) to six (macrophage) independent experiments performed with cells from different donors). Intracellular concentrations were estimated on the basis of a cell volume of 0.2 pl per cell for PBMCs and monocytes and 1 pl per cell for macrophages. **c**, Intracellular NTP levels required for inhibition of EBOV replication in cell culture. The mixture of GS-5734 and its diastereomer on phosphorous was incubated continuously for 72 h at 1 μM and levels of intracellular NTP were determined (results are the average of duplicate incubations performed in each cell type; two independent studies were performed in HMVEC isolated from different donors). The corresponding EBOV EC_{50} values for the prodrug diastereomeric mixture were 100, 184, and 121 nM in human macrophages, HeLa, and HMVEC, respectively, suggesting that an average intracellular NTP concentration of approximately 5 μM is required for 50% inhibition of EBOV *in vitro*.



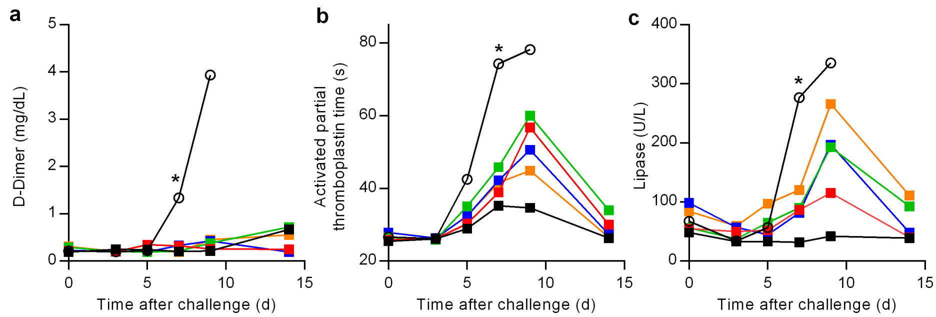
Extended Data Figure 2 | Inhibition of EBOV Makona replication by GS-5734. Huh-7 cells infected with wild-type EBOV (Makona) were incubated for 3 days in the presence of serial dilutions of GS-5734. The amount of infectious virus produced was quantified by endpoint dilution assay of culture media on fresh Vero cell monolayers and the tissue culture infectious dose that caused 50% infection ($TCID_{50}$) was determined. Independently, total RNA was extracted from infected cells and EBOV RNA levels were quantified using a nucleoprotein (NP) gene-specific qRT–PCR. Values represent mean ± s.d. of $\log_{10}$-transformed values, $n = 4$ replicates.



Extended Data Figure 3 | Homology model of RSV A2 (cyan) and EBOV (coral) polymerase based on HIV reverse transcriptase (PDB: 1RTD) with NTP (green and red representing the nucleoside and triphosphate portion, respectively).



Extended Data Figure 4 | Clinical signs of disease in individual rhesus monkeys exposed to Ebola virus. Animals were observed multiple times each day and were subjectively assigned a clinical disease score ranging from 0 to 5 based on responsiveness, posture, and activity. Maximum daily scores were converted to colour code, with darker colours indicative of more severe disease signs. The schematic was truncated to emphasize clinical scores during the acute disease phase, and none of the animals exhibited clinical disease signs outside of the times that are shown.



Extended Data Figure 5 | Amelioration of EVD clinical pathology by GS-5734 in rhesus monkeys. a–c, Group mean ($n = 6$ per group) values of D-dimer (**a**), activated partial thromboplastin time (**b**), and lipase (**c**). Black (open symbols), vehicle; red, 3 mg kg^{-1} d0; green, 3 mg kg^{-1} d2; blue, 10/3 mg kg^{-1} d2; orange, 10/3 mg kg^{-1} d3; black (closed symbols), 10 mg kg^{-1} d3. Error bars omitted for clarity; *x* axes truncated at day 15. *$P < 0.05$ for comparison of mean change from d0 of vehicle and 10 mg kg^{-1} d3 groups at day 7.

Extended Data Table 1 | *In vitro* cytotoxicity of GS-5734 and Nuc in human cell lines and primary cells

	CC_{50} (μM)*		
	GS-5734	Nuc	Puromycin
Human cell lines			
HEp-2	6.0 ± 1.5	> 100	0.53 ± 0.10
HepG2	3.7 ± 0.2	> 100	0.73 ± 0.01
PC-3	8.9 ± 1.6	> 100	0.52 ± 0.11
MT-4	1.7 ± 0.4	69.3 ± 25.7	0.12 ± 0.03
Human primary cells			
Hepatocytes	2.5 ± 0.6	> 100	1.5 ± 0.6
Renal proximal tubular epithelial cells (RPTEC)	12.9 ± 6.2	> 100	1.1 ± 0.3
Quiescent PBMCs	> 20	> 100	6.8 ± 1.4
Stimulated PBMCs	14.8 ± 5.8	> 100	1.6 ± 0.2

*Drug concentrations reducing the cell viability by 50% (CC_{50}) are presented. All CC_{50} values represent the mean ± s.d. of at least two independent experiments. Puromycin was included in experiments as a positive control for cytotoxicity.

Extended Data Table 2 | Individual plasma viral RNA ($\log_{10}$(copies ml^{-1}))

Treatment Description	Animal #	Plasma Viral RNA on Study Day														
		0	2	3	4	5	6	7	8	9	10	12	14	18	21/22	28/29
Vehicle																
	1	ND	ND	ND	6.6	9.0	–	10.0	–	9.5						
	2	ND	ND	ND	5.9	8.9	–	9.8								
	3	ND	ND	6.5	8.2	8.5	–	8.6								
	4	ND	ND	DET	6.6	8.4	–	8.4	–	8.1						
	5	ND	ND	5.8	8.8	10.0	10.3									
	6	ND	ND	5.4	7.4	9.4	–	9.2	8.7							
GS-5734 3 mg/kg D0																
	1	ND	ND	ND	DET	6.0	–	7.3	–	7.4	7.3					
	2	ND	ND	ND	ND	4.9	–	6.8	–	9.9						
	3	ND	ND	ND	DET	5.5	–	6.5	–	5.8	–	DET	ND	–	ND	ND
	4	ND	ND	ND	ND	4.9	–	5.8	–	5.6	–	DET	ND	–	ND	ND
	5	ND	ND	ND	DET	5.7	–	9.1	–	9.1	8.6					
	6	ND	ND	DET	7.2	9.3	–									
GS-5734 3 mg/kg D2																
	1	ND	ND	ND	DET	6.0	–	6.6	–	6.8	–	6.9	6.0	–	ND	ND
	2	ND	ND	ND	6.4	7.6	–	7.1	–	6.8	–	DET	ND	ND		
	3	ND	ND	ND	DET	6.6	–	7.5	–	8.2	–	5.2	ND	–	ND	ND
	4	ND	ND	ND	5.1	7.0	–	7.4	–	6.6	–					
	5	ND	ND	4.9	6.9	8.1	–	8.1	–	6.6	–	5.2	ND	–	ND	ND
	6	ND	ND	ND	DET	5.5	–	6.7	–	5.9	–	ND	ND	–	ND	ND
GS-5734 10/3 mg/kg D2																
	1	ND	DET	5.2	5.8	7.9	–	8.2	–	8.0						
	2	ND	ND	ND	ND	DET	–	DET	–	ND	–	ND	ND	–	ND	ND
	3	ND	ND	ND	DET	6.1	–	7.8	–	7.5						
	4	ND	ND	ND	4.9	5.5	–	6.1	–	DET	–	ND	ND	–	ND	ND
	5	ND	ND	ND	ND	DET	–	8.2	–	8.2						
	6	ND	ND	ND	5.3	7.1	–	6.9	8.1							
GS-5734 10/3 mg/kg D3																
	1	ND	ND	DET	5.4	6.5	–	6.5	–	5.0	–	ND	ND	–	ND	ND
	2	ND	ND	5.3	6.2	7.1	–	6.8	–	6.0	–	ND	ND	–	ND	ND
	3	ND	ND	DET	5.1	6.9	–	7.0	–	7.0	–	6.7	ND	–	ND	ND
	4	ND	ND	ND	DET	7.0	–	8.1	–	8.3	–	6.2	DET	–	ND	ND
	5	ND	ND	ND	ND	DET	–	ND	–	5.5	–	ND	ND	–	ND	ND
	6	ND	ND	DET	5.8	6.6	–	6.2	–	6.8	–	ND	ND	–	ND	ND
GS-5734 10 mg/kg D3																
	1	ND	ND	DET	5.7	6.1	–	6.0	–	DET	–	ND	ND	–	ND	ND
	2	ND	ND	ND	DET	DET	–	DET	–	5.6	–	ND	ND	–	ND	ND
	3	ND	ND	ND	6.7	DET	–	ND	–	ND	–	ND	ND	–	ND	ND
	4	ND	ND	ND	ND	DET	–	ND	–	ND	–	ND	ND	–	ND	ND
	5	ND	ND	ND	ND	ND	–	ND	–	ND	–	ND	ND	–	ND	ND
	6	ND	ND	DET	5.6	6.9	–	5.5	–	5.9	–	ND	ND	–	ND	ND

–, sample not collected (days 6, 8, 10, and 18 were unscheduled samplings of succumbed animals only). DET, detectable, but below the lower limit of quantitation (8.0×10^4 copies ml^{-1}); ND, not detected, that is, below limit of detection.

Extended Data Table 3 | Summary and statistical analysis of plasma viral RNA

	Plasma Viral RNA, mean log_{10} copies ml^{-1} (P value*)					
Day	Vehicle	GS-5734 3 mg/kg D0	GS-5734 3 mg/kg D2	GS-5734 10/3 mg/kg D2	GS-5734 10/3 mg/kg D3	GS-5734 10 mg/kg D3
3	4.77	3.32 (**0.019**)	3.32 (**0.020**)	3.36 (**0.023**)	4.33 (0.454)	3.63 (0.062)
4	7.24	4.66 (**0.001**)	5.52 (**0.024**)	4.49 (**0.001**)	5.06 (**0.005**)	4.81 (**0.002**)
5	9.05	6.04 (**<0.001**)	6.82 (**0.002**)	6.07 (**<0.001**)	6.52 (**0.001**)	5.12 (**<0.001**)
7	9.19	7.09 (**0.013**)	7.24 (**0.015**)	7.00 (**0.007**)	6.28 (**0.001**)	4.24 (**<0.001**)
9	8.76	7.55 (0.351)	6.82 (0.132)	6.30 (0.065)	6.42 (0.072)	4.22 (**0.001**)
12	–	4.90 (NA)	5.05 (NA)	3.00 (NA)	4.14 (NA)	3.00 (NA)

*P values are from ANOVA comparing each GS-5734 treatment group with vehicle group using Dunnett's test to adjust for multiple comparisons ($n = 6$ animals per group, PCR sample assays performed in triplicate). EBOV RNA values reported as '<LOD' were substituted as 10^3 RNA copies ml^{-1}, and values reported as '>LOD, <LLOQ' were substituted as LLOQ of 8.0×10^4 RNA copies ml^{-1} for computation purpose. Statistically significant P values ($P < 0.05$) are highlighted in bold. NA, not applicable, owing to no survivors in vehicle group.

Extended Data Table 4 | Statistical summary of selected clinical pathology parameters

Parameter	Vehicle	Mean Change from Baseline, Day 7 (P value compared with combined vehicle group*) GS-5734 3 mg/kg D0	GS-5734 3 mg/kg D2	GS-5734 10/3 mg/kg D2	GS-5734 10/3 mg/kg D3	GS-5734 10 mg/kg D3
Platelet count (10^3/ µL)	-279	-118 (**0.012**)	-202 (0.055)	-155 (0.055)	-98 (**0.014**)	-65 (**0.008**)
PT (sec)	5.0	1.3 (**0.01**)	3.2 (0.27)	1.6 (0.06)	2.5 (**0.02**)	1.7 (**0.01**)
APTT (sec)	47.7	12.6 (**0.012**)	19.3 (**0.014**)	14.3 (**0.008**)	15.2 (**0.008**)	9.8 (**0.008**)
Fibrinogen (mg/dL)	2.5	-4.7 (**0.012**)	-5.5 (**0.008**)	-5.0 (**0.014**)	-5.0 (**0.014**)	-4.4 (**0.008**)
TT (sec)	50.2	-1.4 (**0.012**)	2.6 (**0.008**)	1.4 (**0.008**)	3.6 (**0.008**)	-1.5 (**0.008**)
Antithrombin (%)	-39.6	-6.1 (**0.012**)	-10.3 (**0.008**)	-7.9 (**0.008**)	5.6 (**0.008**)	3.1 (**0.008**)
D-dimer (mg/dL)	1.15	0.13 (**0.012**)	-0.09 (**0.008**)	0.11 (**0.008**)	-0.12 (**0.005**)	0.02 (**0.007**)
ALT (U/L)	340	14 (**0.012**)	24 (**0.008**)	116 (0.083)	28 (**0.008**)	32 (**0.008**)
AST (U/L)	1425	273 (**0.014**)	206 (**0.014**)	90 (**0.020**)	157 (**0.014**)	36 (**0.014**)
ALP (U/L)	1238	-69 (**0.012**)	-74 (**0.008**)	7 (**0.008**)	8 (**0.008**)	-66 (**0.008**)
CRK (U/L)	5420	1277 (**0.020**)	1002 (**0.014**)	841 (**0.014**)	682 (**0.014**)	96 (**0.014**)
GGT (U/L)	146	-12 (**0.012**)	-13 (**0.008**)	-2 (**0.008**)	1 (**0.008**)	-12 (**0.008**)
LDH (U/L)	8391	1006 (**0.020**)	2263 (**0.014**)	2358 (**0.014**)	2439 (**0.014**)	352 (**0.014**)
Bilirubin (mg/dL)	1.3	0 (0.071)	0 (**0.048**)	0 (**0.048**)	0 (**0.048**)	0 (**0.048**)
BUN (mg/dL)	60	0 (**0.021**)	1 (**0.028**)	2 (**0.036**)	5 (0.055)	1 (**0.021**)
Creatinine (mg/dL)	1.80	0.12 (**0.015**)	0.27 (**0.017**)	0.18 (**0.014**)	0.27 (**0.014**)	0.43 (0.066)
Lipase (U/L)	205	17 (0.14)	34 (0.12)	-17 (0.055)	36 (0.12)	-12 (**0.036**)
Triglycerides (mg/dL)	538	-7 (**0.012**)	52 (**0.008**)	63 (**0.008**)	420 (0.083)	-6 (**0.008**)
CRP (mg/dL)	48.6	48.8 (0.83)	43.8 (1.0)	41.2 (0.31)	35.3 (0.24)	13.2 (**0.008**)
Albumin (g/dL)	-1.5	-0.8 (**0.012**)	-1.2 (0.170)	-0.8 (**0.036**)	-0.8 (**0.022**)	-0.7 (**0.008**)
Total protein (mg/dL)	-1.1	-0.5 (**0.034**)	-0.9 (0.27)	-0.6 (0.17)	-0.4 (**0.035**)	-0.4 (**0.008**)
Chloride (mEq/dL)	-14	-3 (**0.011**)	-5 (**0.008**)	-6 (**0.013**)	-6 (0.067)	0 (**0.008**)
Phosphate (mEq/dL)	0.2	-2.1 (**0.036**)	-2.5 (**0.021**)	-1.5 (0.12)	-1.0 (0.65)	-0.3 (0.93)
Sodium (mg/dL)	-17	-8 (**0.019**)	-10 (**0.042**)	-9 (0.054)	-7 (**0.042**)	-5 (**0.014**)

*Wilcoxon rank-sum test without adjustment for multiple comparisons using a combined vehicle group as a control group for the analysis ($n = 6$ animals per group).
Statistically significant *P* values ($P < 0.05$) are highlighted in bold.
ALP, alkaline phosphatase; ALT, alanine aminotransferase; APTT, activated partial thromboplastin time; AST, aspartate aminotransferase; BUN, blood urea nitrogen; CRK, creatine kinase; CRP, C-reactive protein; GGT, gamma glutamyl transferase; LDH, lactate dehydrogenase; PT, prothrombin time; TT, thrombin time.

Extended Data Table 5 | *L* gene deep sequencing screening sample description and metrics

Treatment Description	Animal #	Day	Survival Outcome	Genbank Accession Number	% *L* Gene Coverage	Polymerase Amino Acid and Codon Changes	% of Population with Change	iSNV Description
Vehicle	6	7	Deceased	KU321182	100.0	W (TGG) @191 (T+G)	21.9	Volatile sub clonal substitution observed in most samples in all treatment groups and controls >2% of population. Causes a frameshifting insertion at the end of a large homopolymer.
3 mg/kg D0	2	7	Deceased	KU321189	100.0	E (GAA) @2173 E (GAg)	99.5	Synonymous substitution unlikely selection pressure.
3 mg/kg D0	6	5	Deceased	KU321084	100.0	G (GGT) @1160 G (GGc)	44.5	Synonymous substitution unlikely selection pressure.
3 mg/kg D2	1	7	Survived	KU321152	100.0	W (TGG) @191 (T+G)	21.8	Volatile sub clonal substitution observed in most samples in all treatment groups and controls >2% of population. Causes a frameshifting insertion at the end of a large homopolymer.
						Q (CAA) @805 (+AA)	26.6	Volatile sub clonal substitution observed in most samples in all treatment groups and controls >2% of population. Causes a frameshifting insertion at the end of a large homopolymer.
3 mg/kg D2	1	9	Survived	KU321162	100.0	W (TGG) @191 (T+G)	13.8	Volatile sub clonal substitution observed in most samples in all treatment groups and controls > 2% of population. Causes a frameshifting insertion at the end of a large homopolymer.
3 mg/kg D2	3	5	Survived	KU321165	100.0	W (TGG) @191 (T+G)	23.4	Volatile sub clonal substitution observed in most samples in all treatment groups and controls >2% of population. Causes a frameshifting insertion at the end of a large homopolymer.
						Q (CAA) @1755 Q (CAG)	99.0	Synonymous mutation unlikely selection pressure.
3 mg/kg D2	4	9	Deceased	KU321088	100.0	W (TGG) @191 (T+G)	13.9	Volatile sub clonal substitution observed in most samples in all treatment groups and controls >2% of population. Causes a frameshifting insertion at the end of a large homopolymer.
10/3 mg/kg D2	3	5	Deceased	KU321098	81.2	—	24.2	Non-coding substitution, unlikely selection pressure.
10/3 mg/kg D3	2	9	Survived	KU321149	100.0	K (AAG) @341 K (AAa)	26.7	Synonymous substitution unlikely selection pressure.
						I (ATC) @348 S (AgC)	22.3	Non-synonymous substitution tolerated by survivor.
10/3 mg/kg D3	3	7	Survived	KU321172	99.3	K (AAA) @1387 K (AAg)	28.3	Synonymous substitution unlikely selection pressure.
						F (TTT) @1827 (TT-)	20.6	Indel causes a frameshifting deletion at the end of a large homopolymer region.
10 mg/kg D3	1	5	Survived	KU321154	86.5	K (AAG) @659 N (AAt)	28.0	Non-synonymous substitution tolerated by survivor.
						Q (CAA) @805 (+AA)	40.2	Volatile sub clonal substitution observed in most samples in all treatment groups and controls >2% of population. Causes a frameshifting insertion at the end of a large homopolymer.

Genomic sequence analysis was conducted on all samples containing quantifiable concentrations of viral RNA (that is, >LLOQ as assessed by quantitative real-time PCR).
Sequences for which no change (defined as >2% of the population) was noted from the reference sequence are not shown.
iSNV, intra-host Single Nucleotide Variant.

doi:10.1038/nature17170

Hepatitis B virus X protein identifies the Smc5/6 complex as a host restriction factor

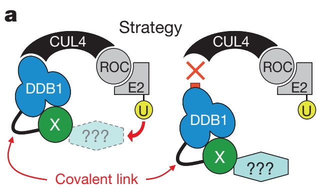
Adrien Decorsière[1]*, Henrik Mueller[1]†*, Pieter C. van Breugel[1]†*, Fabien Abdul[1]*, Laetitia Gerossier[2], Rudolf K. Beran[3], Christine M. Livingston[3], Congrong Niu[3], Simon P. Fletcher[3], Olivier Hantz[2] & Michel Strubin[1]

Chronic hepatitis B virus infection is a leading cause of cirrhosis and liver cancer[1,2]. Hepatitis B virus encodes the regulatory HBx protein whose primary role is to promote transcription of the viral genome, which persists as an extrachromosomal DNA circle in infected cells[3–5]. HBx accomplishes this task by an unusual mechanism, enhancing transcription only from extrachromosomal DNA templates[6]. Here we show that HBx achieves this by hijacking the cellular DDB1-containing E3 ubiquitin ligase to target the 'structural maintenance of chromosomes' (Smc) complex Smc5/6 for degradation. Blocking this event inhibits the stimulatory effect of HBx both on extrachromosomal reporter genes and on hepatitis B virus transcription. Conversely, silencing the Smc5/6 complex enhances extrachromosomal reporter gene transcription in the absence of HBx, restores replication of an HBx-deficient hepatitis B virus, and rescues wild-type hepatitis B virus in a DDB1-knockdown background. The Smc5/6 complex associates with extrachromosomal reporters and the hepatitis B virus genome, suggesting a direct mechanism of transcriptional inhibition. These results uncover a novel role for the Smc5/6 complex as a restriction factor selectively blocking extrachromosomal DNA transcription. By destroying this complex, HBx relieves the inhibition to allow productive hepatitis B virus gene expression.

HBx stimulates transcription of the hepatitis B virus (HBV) genome as well as of any reporter gene provided the DNA template remains extrachromosomal. When the template is integrated into the chromosome, the HBx effect is lost[4,6]. Previous studies suggested that this unusual property requires HBx to bind the DDB1 subunit of the DDB1-containing E3 ubiquitin ligase to target an unknown host cell factor for ubiquitin-mediated degradation[6–9]. To identify the HBx target(s), we employed a tandem affinity purification (TAP) strategy. As baits, we used wild-type HBx, a DDB1-binding-defective HBx(R96E) mutant[8], the woodchuck hepatitis virus HBx counterpart, WHx, that exhibits similar properties[8,10], and the unrelated paramyxovirus SV5-V protein. SV5-V binds DDB1 in a manner similar to HBx to target STAT1 for ubiquitin-mediated degradation[8,11,12]. DDB1 was the most abundant protein species co-purifying with all viral proteins, except the HBx(R96E) mutant (Extended Data Fig. 1a, b). Other components of the E3 ligase were also identified. This provides further evidence that HBx predominantly exists in association with the DDB1 E3 ligase complex[8,13,14]. However, no protein behaving as a potential HBx substrate was recovered. Similarly, STAT1 did not co-purify with SV5-V under these conditions (Extended Data Fig. 1b).

We therefore turned to an alternative strategy. We postulated that HBx may detectably interact with its substrate only when bound to DDB1 and when substrate ubiquitination is blocked. With this in mind, we took advantage of our previous finding that a covalent link between HBx and DDB1 forces interaction between the two proteins, thereby preventing HBx binding to endogenous DDB1[8,13]. We fused TAP-tagged HBx to wild-type DDB1 or to a DDB1 mutant that could not incorporate into the E3 ligase complex (Fig. 1a). The expectation was that fusion of HBx to the DDB1 mutant would prevent ubiquitination of the substrate and thereby stabilize the HBx–substrate interaction. A proof-of-principle experiment demonstrated that this did indeed occur with SV5-V and its target STAT1 (Extended Data Fig. 1c, d). Intriguingly, all six subunits of the Smc5/6 complex (Extended Data Fig. 1e), which plays a broad role in DNA repair and chromatin organization[15–17], were only recovered with the HBx–DDB1 mutant fusion (Fig. 1b and Supplementary Table 1). The selective recovery of this complex with the HBx–DDB1 mutant, and its lack of interaction with the corresponding SV5-V–DDB1 construct, was confirmed by co-immunoprecipitation experiments (Fig. 1c). Hence, the Smc5/6

a



b

HBx–DDB1	WT Pep.	WT Cov. (%)	Mutant Pep.	Mutant Cov. (%)
HBx				
HBx	2	18	3	28
DDB1 E3 ligase				
DDB1	39	38	34	35
DDA1	3	28	5	39
CUL4A	2	4	–	–
SMC complex				
Smc5	–	–	26	25
Smc6	–	–	17	22
Nse1	–	–	3	12
Nse2	–	–	7	23
Nse3	–	–	11	45
Nse4	–	–	2	6

c

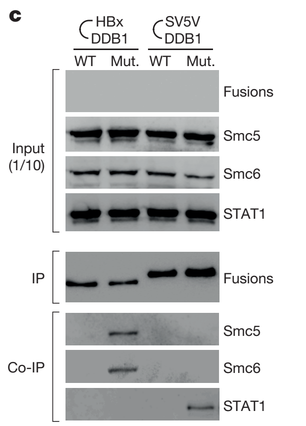


Figure 1 | Identification of the Smc5/6 complex as an HBx interacting partner. a, Strategy. See text for description. **b**, The HBx–DDB1 fusions and associated proteins were isolated by TAP and subjected to mass spectrometric analysis (see Methods). The number of unique peptides (Pep.) and percentage of total protein sequence covered (Cov. (%)) are indicated. Nse1–4 are subunits of the Smc5/6 complex. WT, wild type. **c**, HBx and SV5-V fused to the indicated DDB1 were purified from transfected HepG2 cells. The amount recovered (IP) and the presence of Smc5, Smc6 and STAT1 (Co-IP) were assessed by western blot. The fusion proteins were detected using anti-Flag antibodies; expression was too low for detection in crude extracts (input). For gel source data, see Supplementary Fig. 1.

[1]Department of Microbiology and Molecular Medicine, University Medical Centre (C.M.U.), Rue Michel-Servet 1, 1211 Geneva 4, Switzerland. [2]CRCL, INSERM U1052, CNRS 5286, Université de Lyon, 151, Cours A Thomas, 69424 Lyon Cedex, France. [3]Gilead Sciences, Inc., 333 Lakeside Drive, Foster City, California 94404, USA. †Present addresses: Roche Pharmaceutical Research & Early Development, Infectious Diseases Discovery & Translational Area, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland (H.M.); The Netherlands Cancer Institute, Division of Biological Stress Response, Plesmanlaan 121, 1066CX, Amsterdam, The Netherlands (P.C.v.B.).
*These authors contributed equally to this work.

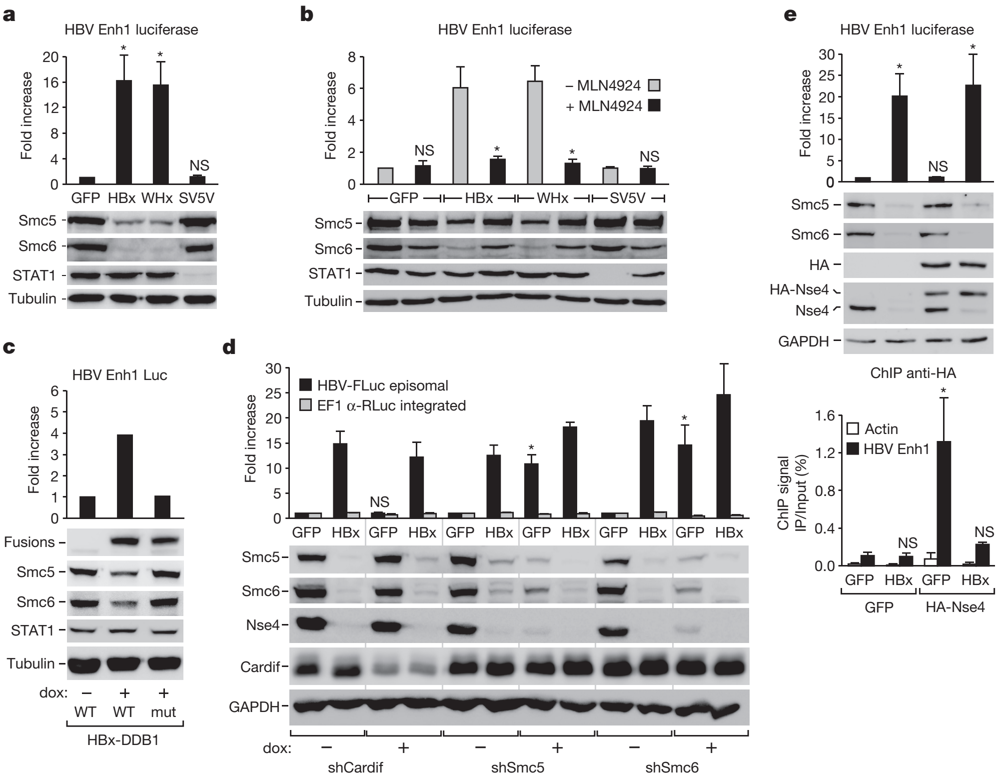


Figure 2 | HBx stimulates reporter gene activity by degrading Smc5/6 to prevent its binding to episomal DNA. a, HepG2 cells were transfected with a luciferase reporter gene driven by the HBV Enhancer I (Enh1) and then transduced with lentiviral vectors expressing green fluorescent protein (GFP) or the indicated GFP-tagged viral proteins. Luciferase assay and western blot were performed 6 days after transfection. Luciferase activity is relative to GFP, which was set to 1. *$P < 0.05$ by paired analysis of variance (ANOVA) relative to GFP control. NS, not significant ($P \geq 0.05$). **b**, Same as in **a** except cells were treated or not with 5 μM MLN4924 starting 6 h before transduction. Analysis was performed 3 days later. *$P < 0.05$ by t-test relative to matched no MLN4924 control. **c**, Same as in **a** but with cells expressing a doxycycline (dox)-inducible wild-type or mutant (mut) HBx–DDB1 fusion. Analysis performed as in **a**. See also Extended Data Fig. 2e. **d**, Activity of a transiently transfected (episomal) HBV Enh1-driven firefly reporter (HBV-FLuc) and a chromosomally integrated EF1α promoter-driven *Renilla* reporter (EF1α-RLuc) in HepG2 cells transduced with GFP or GFP–HBx and expressing (+) or not (−) the indicated doxycycline-inducible shRNA. Analysis performed after 6 days of culture. *$P < 0.05$ by t-test relative to matched HBV-FLuc episomal no doxycycline GFP control. **e**, Same as in **a** but with HepG2 cells co-transduced with GFP or an HA-tagged Nse4 subunit of the Smc5/6 complex together with GFP or GFP–HBx as indicated. Luciferase and western blot analyses performed 6 days later. In parallel, HA-Nse4 binding to the episomal reporter (filled bars) or the β-actin gene (open bars) was monitored by ChIP using anti-HA antibodies (bottom). Data expressed as percentage of input DNA recovered. Note the reduced binding of HA-Nse4 in the presence of HBx despite comparable expression levels, consistent with the protein associating to DNA only when incorporated into the Smc5/6 complex. See also Extended Data Fig. 3c. *$P < 0.05$ by paired ANOVA relative to GFP control. Data in **a**, **b**, **d** and **e** are mean ± s.e.m.; $n = 3$ independent experiments.

complex behaves as STAT1 in the SV5-V experiment, strongly suggesting that the complex is a substrate of HBx.

This hypothesis and the key role of the Smc5/6 complex in mediating the HBx stimulatory activity were confirmed as follows. Western blot analysis revealed that Smc5 and Smc6 levels were strongly decreased in HepG2 cells expressing HBx (Fig. 2a). The same was observed with the woodchuck WHx, suggesting a conserved mechanism (Fig. 2a). This correlated with the stimulatory effects of HBx and WHx on the activity of a transiently transfected luciferase reporter construct, which, like the HBV genome, remains extrachromosomal[6]. In contrast, SV5-V triggered degradation of STAT1 but had no effect on Smc5 or Smc6 levels or reporter gene expression (Fig. 2a). HBx-mediated depletion of Smc6 was rapid, occurring concomitantly with or preceding the HBx-mediated increase in reporter gene expression (Extended Data Fig. 2a), and was inhibited by exposure to the proteasome inhibitor MG132 (Extended Data Fig. 2b), consistent with a role of proteasomes in HBx activity[18]. HBx and WHx had no impact on Smc5 and Smc6 messenger RNA levels, excluding an effect on transcription or mRNA stability (Extended Data Fig. 2c). Furthermore, pre-treatment of cells with MLN4924, an inhibitor of the E3 ubiquitin ligase family to which the DDB1 E3 ligase belongs[19,20], abrogated the effects of HBx and WHx both on reporter gene activity and on Smc5 and Smc6 protein levels (Fig. 2b and Extended Data Fig. 2d). This implicates the DDB1 E3 ubiquitin ligase activity in HBx function. Consistently, HBx retained transactivation activity and the ability to decrease Smc5 and Smc6 protein levels when fused to wild-type DDB1 but not to the mutant that cannot incorporate into the E3 ligase (Fig. 2c and Extended Data Fig. 2e).

The above results are consistent with HBx and WHx using the DDB1 E3 ubiquitin ligase to target the Smc5/6 complex for ubiquitin-mediated proteasome degradation. They further suggest that this event explains the stimulatory effect of HBx on extrachromosomal reporter gene activity. To test the latter hypothesis, we monitored the consequence of depleting Smc5 or Smc6 by RNA interference. HepG2 cell lines were generated to express doxycycline-inducible short hairpin (sh) RNA targeting Smc5, Smc6, or Cardif (MAVS) as a control. Dual *Renilla* and firefly luciferase reporter genes driven by unrelated promoters were tested in pairwise combinations in these cells, with one being integrated into the chromosomes and the other transiently transfected, and therefore episomal. All three reporter constructs tested responded to HBx only as episomal templates (Fig. 2d and Extended Data Fig. 2f, g). This was consistent with our earlier observation that HBx stimulates extrachromosomal DNA templates selectively and independently of promoter type[6]. Depletion of Smc5 or Smc6, which destabilizes the

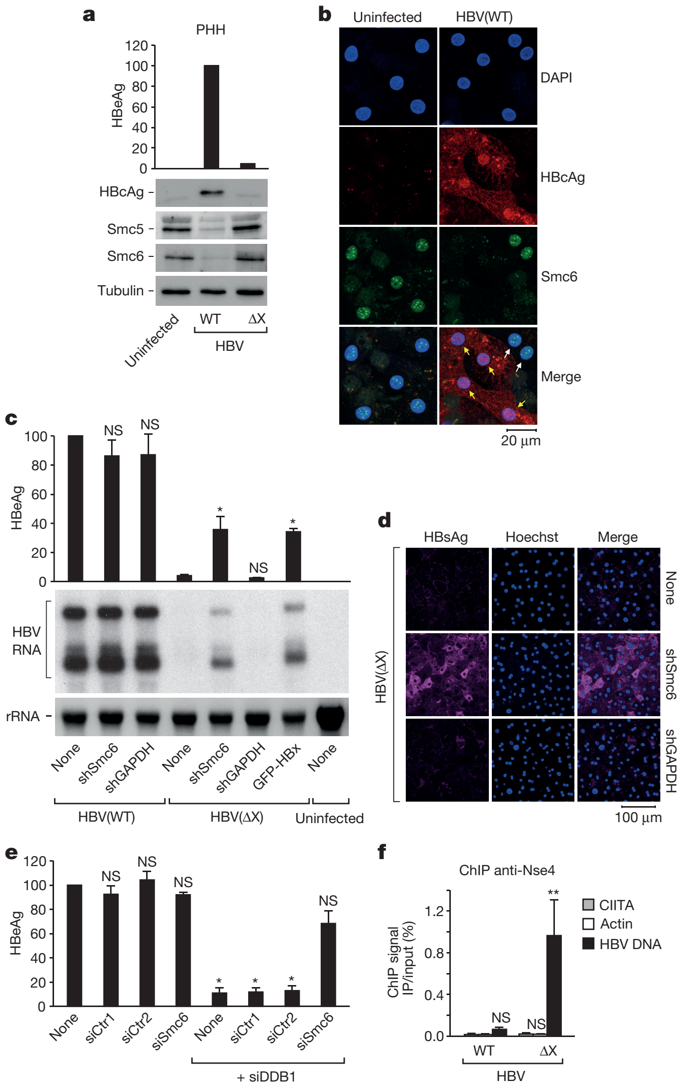


Figure 3 | HBx promotes HBV transcription in PHHs by preventing Smc5/6 binding to the viral genome. a, Purified PHHs were left uninfected or infected with wild-type or HBx-deficient (ΔX) HBV. HBeAg secretion, a marker for HBV gene expression, and indicated proteins were quantified 10 days later by ELISA and western blot. HBcAg is the viral capsid. HBeAg concentrations are relative to wild-type HBV, which was set to 100. See also Extended Data Fig. 6a. **b**, Confocal microscopy of control (left) and HBV-infected (right) PHHs stained for HBcAg (red) and Smc6 (green). Cell nuclei visualized with 4′,6-diamidino-2-phenylindole (DAPI, blue). HBcAg-positive and -negative cells are indicated by yellow and white arrows, respectively. See also Extended Data Fig. 7. **c**, PHHs were mock transduced (none) or transduced 1 day before HBV infection with indicated shRNAs or GFP–HBx. HBeAg secretion and viral RNA production were assessed 8 days later by ELISA and northern blot. Mean ± s.e.m.; $n = 4$ independent experiments with three different PHH donors. $*P < 0.05$ by paired ANOVA relative to matched no shRNA control (none). **d**, Same as in **c** but cells infected with HBx-mutant HBV were examined for surface antigen (HBsAg) positivity 12 days after infection. Cell nuclei stained with Hoechst dye (blue). **e**, PHHs were mock transfected (none) or transfected with indicated siRNAs 4 days before HBV infection. HBeAg secretion was measured 13 days later. Mean ± s.e.m.; $n = 4$ independent experiments with two different PHH donors. $*P < 0.05$ by paired ANOVA relative to no siRNA control (none). **f**, PHHs were infected as indicated. At 10 days after infection, the binding of Smc5/6 to repressed CIITA and active β-actin genes, and to episomal HBV DNA was monitored by ChIP using anti-Nse4 antibodies. Mean ± s.e.m.; $n = 3$ independent experiments with different PHH donors and virus stocks. $**P < 0.01$ by unpaired ANOVA relative to WT HBV β-actin control.

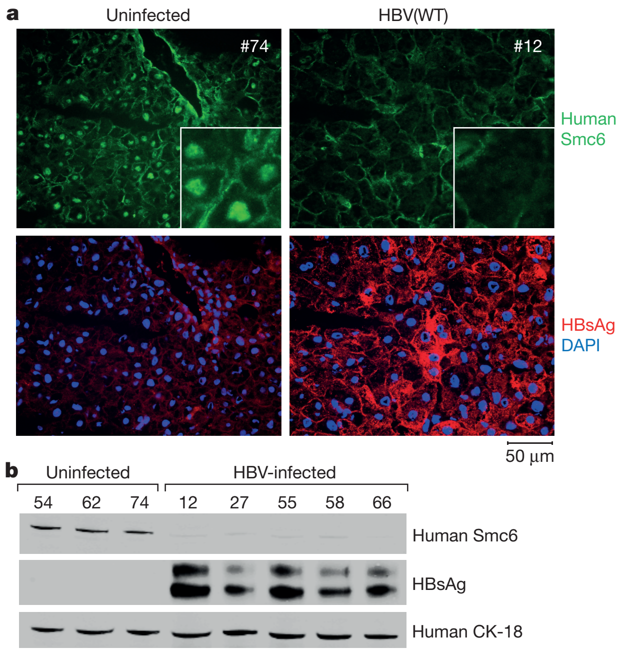


Figure 4 | HBV infection induces Smc6 degradation in humanized mouse liver tissue. a, Fresh-frozen liver tissue sections from the indicated humanized mice were stained for human Smc6 (green) and HBsAg (red). Blue, DAPI. Insets represent a digital zoom of selected areas. Note the absence of Smc6 nuclear staining in the HBV-infected tissue. See also Extended Data Figs 8 and 9. Images are representative of six HBV-infected and three uninfected humanized mice. **b**, Western blot analysis for human Smc6 and HBsAg expression in the indicated humanized liver samples. Human cytokeratin-18 (CK-18) served as a control.

entire Smc5/6 complex[21], stimulated expression of the three episomal constructs to levels approaching those measured in the presence of HBx, but remained without effect on the corresponding chromosomal reporters (Fig. 2d and Extended Data Fig. 2f, g). This recapitulated exactly the situation observed with HBx. Moreover, knockdown of Smc5 or Smc6 had no further effect on transcription stimulated by HBx, consistent with activation by HBx and by Smc5/6 knockdown occurring through the same pathway. HBx-mediated destruction of the Smc5/6 complex, which has been involved in cell cycle progression and recombinational DNA repair[22–24], did not alter the cell cycle or induce integration of the reporter gene into the chromosomes (Extended Data Fig. 3a, b). Instead, chromatin immunoprecipitation (ChIP) analysis revealed that the Smc5/6 complex binds to transiently transfected reporter plasmids but not to a chromosomally integrated reporter, and that HBx reduces this binding (Fig. 2e and Extended Data Fig. 3c). This is in line with recent *in vitro* studies[25]. Thus, HBx stimulates gene transcription mainly, if not exclusively, by triggering degradation of the Smc5/6 complex to prevent its association with episomal DNA templates.

The importance of these findings for HBV replication was examined using an infection assay of primary human hepatocytes (PHHs). In this assay, mutant HBV lacking a functional HBx gene is fully competent to enter cells and deliver its circular DNA genome into the nucleus. However, no viral gene transcription is detected[4] (Fig. 3c and Extended Data Fig. 4a). We observed that cells infected at high efficiency (Extended Data Fig. 5) with wild-type HBV displayed decreased levels of Smc5 and Smc6, but not when infected with an HBx-defective virus (Fig. 3a, b and Extended Data Figs 6a and 7). Conversely, depletion of Smc6 by shRNA (Extended Data Fig. 6b) restored replication competency to the HBx-deficient virus without affecting genome copy number (Extended Data Fig. 4b) but had no effect on replication of the wild-type virus (Fig. 3c, d and Extended Data Fig. 6c). The same was observed in HBV-permissive HepaRG

cells (Extended Data Fig. 6d). Furthermore, siRNA-mediated silencing of DDB1 blocked HBV transcription, and this effect was relieved by the concomitant downregulation of Smc6 or Smc5 (Fig. 3e and Extended Data Fig. 6e–i). Thus, a major function of the HBx–DDB1 E3 ligase interaction is to target the Smc5/6 complex for destruction. ChIP analysis demonstrated that the Smc5/6 complex also associates with the HBV episomal DNA genome, and that binding was reduced in the presence of HBx (Fig. 3f).

In summary, our results indicate that the Smc5/6 complex binds episomal DNA templates and inhibits their transcription. HBx, by destroying the complex, relieves the inhibition to allow productive HBV gene expression. Further supporting this notion, HBV-infected humanized mouse liver tissues show markedly reduced levels of Smc6 (Fig. 4 and Extended Data Figs 8 and 9). This points to a novel role for the Smc5/6 complex as a host restriction factor, and offers a potential new avenue for therapeutic intervention against HBV infection. The Smc5/6 complex is, together with condensin and cohesin, one of the three related Smc complexes identified in eukaryotes[26]. These complexes share a similar architecture and play fundamental roles in chromosome organization, segregation, and repair (reviewed in ref. 27). The Smc5/6 complex has well-recognized functions in homologous recombination-mediated DNA repair[15,16]. In addition, the complex has been implicated in DNA replication, chromosome topology, and transcription, but little is known of its precise function[15–17,24]. A role in host defence against viral infection has not previously been reported. It will be interesting to determine whether the restrictive activity of the Smc5/6 complex is connected to any of its cellular functions, and whether the complex exhibits antiviral activity against other episomal DNA viruses.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 February 2014; accepted 27 January 2016.

1. Seeger, C. & Mason, W. S. Hepatitis B virus biology. *Microbiol. Mol. Biol. Rev.* **64,** 51–68 (2000).
2. Ganem, D. & Prince, A. M. Hepatitis B virus infection — natural history and clinical consequences. *N. Engl. J. Med.* **350,** 1118–1129 (2004).
3. Benhenda, S., Cougot, D., Buendia, M. A. & Neuveut, C. Hepatitis B virus X protein molecular functions and its role in virus life cycle and pathogenesis. *Adv. Cancer Res.* **103,** 75–109 (2009).
4. Lucifora, J. *et al.* Hepatitis B virus X protein is essential to initiate and maintain virus replication after infection. *J. Hepatol.* **55,** 996–1003 (2011).
5. Feitelson, M. A., Bonamassa, B. & Arzumanyan, A. The roles of hepatitis B virus-encoded X protein in virus replication and the pathogenesis of chronic liver disease. *Expert Opin. Ther. Targets* **18,** 293–306 (2014).
6. van Breugel, P. C. *et al.* Hepatitis B virus X protein stimulates gene expression selectively from extrachromosomal DNA templates. *Hepatology* **56,** 2116–2124 (2012).
7. Leupin, O., Bontron, S., Schaeffer, C. & Strubin, M. Hepatitis B virus X protein stimulates viral genome replication via a DDB1-dependent pathway distinct from that leading to cell death. *J. Virol.* **79,** 4238–4245 (2005).
8. Li, T., Robert, E. I., van Breugel, P. C., Strubin, M. & Zheng, N. A promiscuous α-helical motif anchors viral hijackers and substrate receptors to the CUL4-DDB1 ubiquitin ligase machinery. *Nature Struct. Mol. Biol.* **17,** 105–111 (2010).
9. Hodgson, A. J., Hyser, J. M., Keasler, V. V., Cang, Y. & Slagle, B. L. Hepatitis B virus regulatory HBx protein binding to DDB1 is required but is not sufficient for maximal HBV replication. *Virology* **426,** 73–82 (2012).
10. Sitterlin, D., Bergametti, F. & Transy, C. UVDDB p127-binding modulates activities and intracellular distribution of hepatitis B virus X protein. *Oncogene* **19,** 4417–4426 (2000).
11. Ulane, C. M. & Horvath, C. M. Paramyxoviruses SV5 and HPIV2 assemble STAT protein ubiquitin ligase complexes from cellular components. *Virology* **304,** 160–166 (2002).
12. Precious, B., Childs, K., Fitzpatrick-Swallow, V., Goodbourn, S. & Randall, R. E. Simian virus 5 V protein acts as an adaptor, linking DDB1 to STAT2, to facilitate the ubiquitination of STAT1. *J. Virol.* **79,** 13434–13441 (2005).
13. Leupin, O., Bontron, S. & Strubin, M. Hepatitis B virus X protein and simian virus 5 V protein exhibit similar UV-DDB1 binding properties to mediate distinct activities. *J. Virol.* **77,** 6274–6283 (2003).
14. Benhenda, S. *et al.* Methyltransferase PRMT1 is a binding partner of HBx and a negative regulator of hepatitis B virus transcription. *J. Virol.* **87,** 4360–4371 (2013).
15. Murray, J. M. & Carr, A. M. Smc5/6: a link between DNA repair and unidirectional replication? *Nature Rev. Mol. Cell Biol.* **9,** 177–182 (2008).
16. De Piccoli, G., Torres-Rosell, J. & Aragón, L. The unnamed complex: what do we know about Smc5-Smc6? *Chromosome Res.* **17,** 251–263 (2009).
17. Jeppsson, K., Kanno, T., Shirahige, K. & Sjögren, C. The maintenance of chromosome structure: positioning and functioning of SMC complexes. *Nature Rev. Mol. Cell Biol.* **15,** 601–614 (2014).
18. Zhang, Z., Sun, E., Ou, J. H. & Liang, T. J. Inhibition of cellular proteasome activities mediates HBX-independent hepatitis B virus replication *in vivo*. *J. Virol*. **84,** 9326–9331 (2010).
19. Soucy, T. A. *et al.* An inhibitor of NEDD8-activating enzyme as a new approach to treat cancer. *Nature* **458,** 732–736 (2009).
20. Emanuele, M. J. *et al.* Global identification of modular cullin-RING ligase substrates. *Cell* **147,** 459–474 (2011).
21. Taylor, E. M., Copsey, A. C., Hudson, J. J., Vidot, S. & Lehmann, A. R. Identification of the proteins, including MAGEG1, that make up the human SMC5-6 protein complex. *Mol. Cell. Biol.* **28,** 1197–1206 (2008).
22. Kegel, A. & Sjögren, C. The Smc5/6 complex: more than repair? *Cold Spring Harb. Symp. Quant. Biol.* **75,** 179–187 (2010).
23. Potts, P. R., Porteus, M. H. & Yu, H. Human SMC5/6 complex promotes sister chromatid homologous recombination by recruiting the SMC1/3 cohesin complex to double-strand breaks. *EMBO J.* **25,** 3377–3388 (2006).
24. Tapia-Alveal, C., Lin, S. J. & O'Connell, M. J. Functional interplay between cohesin and Smc5/6 complexes. *Chromosoma* **123,** 437–445 (2014).
25. Kanno, T., Berta, D. G. & Sjögren, C. The Smc5/6 complex is an ATP-dependent intermolecular DNA linker. *Cell Reports* **12,** 1471–1482 (2015).
26. Aragon, L., Martinez-Perez, E. & Merkenschlager, M. Condensin, cohesin and the control of chromatin states. *Curr. Opin. Genet. Dev.* **23,** 204–211 (2013).
27. Hirano, T. At the heart of the chromosome: SMC proteins in action. *Nature Rev. Mol. Cell Biol.* **7,** 311–322 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are most grateful to M. Rivoire for providing liver samples, L. Lefrancois and M. Michelet for help in preparing hepatocytes, U. Protzer for the HBV(ΔX)-producing cell line, C. J. Gloeckner for the StrepII/Flag TAP-tag construct, C. E. P. Goldring for the tetracyclin-inducible HepG2 cell line, S. Elledge for the lentiviral pINDUCER vectors, A. R. Lehmann for anti-Smc5 and anti-Smc6 antibodies, M. A. Petit for anti-HBsAg and anti-HBcAg antibodies, D. Garcin for the Cardif shRNA construct, Q. Seguín-Estévez, Y. Grimaldi and P. Ferrari for help with the ChIP assay, P. Arboit and the Geneva Proteomics Core Facility for mass spectrometry analysis, the Centre d'imagerie Quantitative Lyon-Est for help in confocal microscopy, A. Joshi for statistical analysis support, and J. Curran for reading the manuscript. This study was supported by grants from CLARA (Lyon) and the French National Agency for Research against AIDS and viral hepatitis (ANRS) (to O.H.) and from the Swiss National Science Foundation (31003A-127384 and 310030-149626) (to M.S.) and by the Canton of Geneva (to M.S).

Author Contributions A.D. performed the experiments shown in Figs 1c and 2c–e and Extended Data Figs 2e–g and 3b, c, and performed the ChIP assay in Fig. 3f. H.M. performed the experiments shown in Fig. 2a, b and Extended Data Fig. 2a, d. P.C.v.B. established the TAP approach and performed the experiments shown in Fig. 1b and Extended Data Fig. 1. F.A. performed the experiments in Extended Data Figs 2b, c and 3a, c, the western blots in Fig. 3a and Extended Data Fig. 6b, and contributed to Fig. 2e and Extended Data Fig. 3c. O.H. and L.G. performed the PHH infection experiments in Fig. 3a, c, d and Extended Data Figs 4, 5 and 6a–c, the HepaRG experiment in Extended Data Fig. 6d, and contributed to Fig. 3f. C.M.L. performed the confocal and epifluorescence microscopy experiments in Figs 3b and 4a and Extended Data Figs 7–9 and contributed to Fig. 4b. C.N. and R.K.B. performed the PHH experiments in Fig. 3e and Extended Data Fig. 6e, f, h, i. R.K.B. performed the western blots in Fig. 4b and Extended Data Fig. 6g and contributed to Fig. 3f. All authors designed the experiments and interpreted the results. M.S. wrote the paper with input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.H. (olivier.hantz@inserm.fr), S.P.F. (Simon.Fletcher@gilead.com) or M.S. (michel.strubin@unige.ch).

METHODS

Expression plasmids. HBx fused at its carboxy (C) terminus via a 10-amino-acid linker (HMRSRSGLET) either to wild-type DDB1 or to the CUL4-binding-defective DDB1 mutant has been described[8,28]. The SV5-V–DDB1 fusions were generated by replacing the sequence encoding HBx within the HBx–DDB1 fusions by the full-length SV5-V coding region[13]. Unless fused to DDB1, all viral proteins contain an amino (N)-terminal GFP tag and have been described[6]. For TAP, the proteins carried an N-terminal Flag tag followed by two Strep-tag II sequences (FS-tag)[29]. For ChIP analysis, the full-length NSE4a coding region was amplified by PCR from a Namalwa B cell complementary DNA (cDNA) library and fused N-terminal to three tandem copies of the HA epitope. Proteins were expressed from a stably integrated tetracycline-inducible pTRE2hyg plasmid vector (Clontech) in Fig. 1b and Extended Data Fig. 1a, b, from the episomal Epstein–Barr virus-based expression vector EBS-PL[30] in Fig. 1c and Extended Data Fig. 1d, from the tetracycline-inducible lentiviral vector pINDUCER20 (ref. 31) (a gift from S. Elledge) in Fig. 2c and Extended Data Fig. 2e, and from the lentivirus vector pWPT[6,7] in all the other figures.

Luciferase reporter constructs. The luciferase reporter genes were constructed into pcDNA3 (pGL3 in the case of HBV Enhancer I[6]) for transient transfection and into the self-inactivating lentiviral vector pWPT (pWPXL in the case of HBV Enhancer I[6]) for chromosomal integration. In Extended Data Fig. 3c, the EF1α episomal reporter was delivered using an integrase-defective (D116A) lentiviral vector[32].

RNA interference. The short hairpin RNAs (shRNAs) used for Smc5 and Smc6 knockdown were cloned into the miR30 context as 116-nucleotide XhoI–EcoRI fragments into pINDUCER10 (ref. 31) and/or pGIPZ (Open Biosystems) lentiviral vectors. The sequences were as follows, with the hairpin sequence in capital letters and flanking miR30 sequences shaded in bold type: shSmc5(1), **gagcg**GTGAGGTGAAAGAAGTGTTTCTtagtgaagccacagatgtaAGAAACACTTCTTTCACCTCAT**tgcct**; shSmc5(2), **gagcg**GTGCGAAACTTGTTACCGAATTtagtgaagccacagatgtaAATTCGGTAACAAGTTTCGCAT**tgcct**; shSmc6(1), **gagcg**GTGAGCAGCTTTGTAAACGAATtagtgaagccacagatgtaATTCGTTTACAAAGCTGCTCAT**tgcct**; shSmc6(2) (from Thermo Scientific (V3LHS_325916), **gagcg**CAGAGACAGTGCTACTAATCAAtagtgaagccacagatgtaTTGATTAGTAGCACTGTCTCTA**tgcct**.

The constructs used were shSmc5(2) and shSmc6(1) in Fig. 2d, shSmc5(1) and shSmc6(2) in Extended Data Fig. 2f, and shSmc6(2) in Extended Data Fig. 2g. The Cardif shRNA construct in pINDUCER10 was a gift from D. Garcin. The GAPDH shRNA construct in pGIPZ was purchased from Thermo Scientific (RHS4371). Expression was from pINDUCER10 in Fig. 2d and Extended Data Fig. 2f, g and from pGIPZ in Fig. 3c, d. Details of plasmid constructions are available upon request. The sequences of the siRNAs used in the other figures can be found in Supplementary Table 2.

Cell culture and establishment of stable cell lines. The human hepatoma cell lines HepG2 (ATCC HB-8065), HepG2$^{\text{tet-on}}$ (ref. 33), and derivatives were grown at 37 °C under 5% CO_2 in modified Eagle's medium (MEM; Life Technologies or Sigma-Aldrich) supplemented with 10% (vol/vol) fetal calf serum (Gibco), 1% (v/v) penicillin/streptomycin solution, 2 mM L-glutamine, 1 mM sodium pyruvate, and 0.1 mM MEM non-essential amino acids solution (all from Life Technologies). In Extended Data Fig. 6d, HepaRG cells were cultured, induced to differentiate, and infected with HBV as described[34]. The cell lines were not authenticated or tested for *Mycoplasma* contamination because of their direct purchase from ATCC or low passage number.

The stable HepG2 cell lines used in Fig. 1b and Extended Data Fig. 1a, b that expressed the tetracycline-inducible transactivator rtTA and various FS-tagged proteins from a tetracycline-responsive promoter were established as follows. The HepG2$^{\text{tet-on}}$ parental cell line[33] was plated at a density of 1×10^5 cells per 6-cm dish and transfected using FuGENE HD (Roche) with the relevant constructs cloned into pTRE2hyg (Clontech), which carries a hygromycin resistance marker. At day 3 after transfection, cells were selected in medium containing 250 μg ml^{-1} hygromycin B (PAA Laboratories) and 100 μg ml^{-1} G418 (Geneticin, Gibco)[33]. Resistant colonies were picked and expanded in 48-well plates under the same selection conditions. Individual clones were assayed by reverse transcription (RT)-PCR and western blot analyses for transgene expression in the presence or absence of 2 μg ml^{-1} doxycycline (Sigma Aldrich).

The HepG2 cells in Fig. 2c expressing the doxycycline-inducible HBx–DDB1 fusions from pINDUCER20 lentiviral vector carrying a neomycin resistance marker were obtained by selection for 2 days with 1 mg ml^{-1} G418 (Promega) starting at day 3 after transduction.

The HepG2 cells in Fig. 2d and Extended Data Fig. 2f, g expressing doxycycline-inducible shRNAs against Smc5, Smc6, or Cardif from pINDUCER10 lentiviral vector were obtained by selection with 10 μg ml^{-1} puromycin (Calbiochem) for 4 days starting at day 3 after transduction. The resulting puromycin-resistant cells were then cultured for 2 weeks in the absence or presence of 2.5 μg ml^{-1} doxycycline before transduction of GFP or GFP–HBx to maximally induce shRNA expression. Fluorescence microscopy indicated that >90% of the cells were expressing turboRFP (tRFP) from the tRFP-shRNA cassette[31].

The HepG2.2.15 (ref. 4), HepAD38 (ref. 35), and HepG2-H1.3x- (ref. 4) stable cell lines used to produce HBV virions carrying either a wild-type genome or a genome bearing a defective HBx gene have been described. They were maintained in Dulbecco's modified Eagle's medium/F-12 (DMEM/F12) complemented with 10% fetal calf serum, 1% non-essential amino acids, 50 μg ml^{-1} kanamycin, and 200 μg ml^{-1} geneticin.

Transfection, transduction, and reporter gene assay. Transfection of plasmid DNA in HepG2 cells was performed using X-tremeGENE HP DNA Transfection Reagent (Roche) following the manufacturer's instructions. Transfection of siRNA duplexes in HepaRG cells was performed using DharmaFECT-1 (Thermo Scientific) transfection reagent and in PHHs using Lipofectamine RNAiMax (Life Technologies). Recombinant lentivirus production and transduction of HepG2 cells were performed as described previously[7,36]. For PHH transduction, lentivirus supernatants were concentrated by ultracentrifugation at 130,000 *g* for 1 h in a SW32 Beckman rotor at 4 °C and transduction was performed overnight in the presence of 10 ng ml^{-1} of EGF[37]. For luciferase reporter gene assay and western blot analysis, cells were typically seeded at a density of about 6×10^5 cells per 30 mm diameter well (1×10^5 cells per square centimetre) and transfected the next day with 30 ng of reporter plasmid DNA and 2 μg of empty EBS-PL vector. For the ChIP experiment in Fig. 2e, 3 or 30 ng of reporter plasmid were used with similar results.

In Fig. 2d and Extended Data Fig. 2f, g, cells that had been cultured in the absence or presence of doxycyclin for 2 weeks to induce shRNA expression from pINDUCER10 were first transduced and the next day transfected with, respectively, lentiviral and plasmid constructs carrying the *Renilla* and firefly luciferase reporter genes as indicated in the figures. Cells were trypsinized the following day, washed with phosphate-buffered saline (PBS), replated in 6- or 24-well dishes at a density of about 6×10^4 cells per square centimetre and transduced 6 h after replating with lentiviral vectors encoding GFP or the GFP-tagged viral proteins. In Fig. 2b, dimethylsulphoxide (DMSO) or 5 μM MLN4924 (Active Biochem) was added 6 h before transduction and culture medium was replaced daily with fresh medium containing (or not) MLN4924 for 3 days. In Fig. 2c, d and Extended Data Fig. 2f, g, cells were constantly grown in the absence or presence of 2 μg ml^{-1} doxycycline. Cell lysates were prepared 6 days later (3 days in Fig. 2b) for luciferase assay and western blot analysis. Luciferase activities were measured using the Luciferase Assay System (Promega) or Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions. The activities were normalized to protein concentrations as measured by the Bradford assay (Bio-Rad).

Purification of Flag–StrepII-tagged proteins. The two-step purification scheme was adapted from ref. 29. The HepG2$^{\text{tet-on}}$ cell lines conditionally expressing FS-tagged proteins from a doxycycline-inducible promoter were seeded on thirty 15-cm dishes and allowed to reach 90% confluence. Expression of the FS-tagged proteins was then induced by addition of doxycycline (2 μg ml^{-1}) for 48 h. Cells were rinsed once with 20 ml PBS (per plate) and scraped off after incubation on ice for 20 min in 1.5 ml lysis buffer (50 mM HEPES/KOH, pH 7.5, 150 mM NaCl, 5 mM KCl, 5 mM $MgCl_2$, 50 μM $ZnCl_2$, 0.1% IGEPAL and protease inhibitor cocktail from Sigma). The lysates were pooled, supplemented with glycerol (10% final), and clarified by centrifugation at 20,000*g* for 20 min at 4 °C in a Sorvall HB-6 rotor. Supernatants were collected, passed through a 0.45 μm Millipore PVDF filter, and mixed in a 50 ml Falcon tube with a 100 μl packed bead volume of Strep-Tactin Sepharose (IBA) washed with lysis buffer. After incubation for 2 h at 4 °C under constant rotation, the suspension was poured into a 10 ml disposable column (Bio-Rad). After sedimentation, the resin was washed twice with 10 ml lysis buffer containing 10% glycerol. The bound proteins were released by incubating the resin with 10 ml of lysis buffer supplemented with 10% glycerol and 2.5 mM desthiobiotin (IBA 2× Tactin Elution Buffer) for 10 min on ice. The eluted material was directly mixed with a 100 μl packed volume of washed anti-Flag Sepharose beads (Sigma) and incubated with rotation for 1 h at 4 °C in a sealed 10 ml disposable column (Bio-Rad). Subsequently, the column was drained by gravity and the resin washed five times with 1 ml lysis buffer adjusted to 10% glycerol. Bound proteins were recovered by adding five times sequentially 200 μl lysis buffer supplemented with 200 μg ml^{-1} Flag peptide (Sigma) and gently mixing for 5 min on ice. The eluted material from three such purifications was pooled and precipitated with acetone (80% final concentration). After centrifugation at 4 °C in an Eppendorf microcentrifuge, the protein pellet was air-dried, resuspended in 1× Laemmli sample buffer, and run into a 4% SDS–PAGE stacking gel. The protein band was excised and processed for mass spectrometry analysis using an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific). Listed in Fig. 1b are proteins identified

with 99% certainty and represented by at least two peptides ascertained at a 95% confidence level.

In Fig. 1c, extracts from 6×10^6 HepG2 cells seeded on a 10-cm culture dish were prepared 2 days after transfection with 2 μg of EBS-PL expression plasmids. The experiment in Extended Data Fig. 1d was performed in the same way except that the transfected cells were selected in medium containing 200 μg ml^{-1} hygromycin B (PAA Laboratories) and extracts were prepared from a 15-cm dish of confluent cells. Proteins were purified by incubation with 25 μl packed bead volume of Strep-Tactin Sepharose (IBA).

Cell cycle analysis. The cell cycle analysis in Extended Data Fig. 3a was performed starting with 1.5×10^6 HepG2 cells transduced with GFP or GFP–HBx and cultured as above. Cells were harvested, washed once with PBS, and recovered by centrifugation at 500*g* for 5 min. Cells were fixed in 1 ml ice-cold 70% ethanol added dropwise while vortexing and incubated for 30 min on ice. The fixed cells were pelleted by centrifugation at 800*g* for 15 min, washed twice with PBS and directly resuspended in 50 μl PBS supplemented with 100 μg ml^{-1} RNase A (Roche). Cellular DNA was stained by addition of 600 μl of propidium iodide (Sigma Aldrich; 50 μg ml^{-1} in PBS) and incubation for 10 min at room temperature (20–25 °C) and overnight at 4 °C. Samples were analysed using a BD Accuri C6 flow cytometer.

Western blotting. With the exception of Fig. 4b and Extended Data Fig. 6g (see below), western blot analysis was performed as described[38] except that cells were disrupted in 2% SDS in water and protein concentration was estimated using the BCA Protein Assay (Novagen). The membranes were probed with 1:5,000 mouse anti-Flag antibodies (Sigma A2220) to detect the FS-tagged HBx–DDB1 and SV5-V–DDB1 fusion proteins, 1:5,000 mouse monoclonal anti-GFP (Roche 11814460001) to detect GFP–HBx in Extended Data Fig. 2a, b, 1:5,000 mouse monoclonal anti-HA (clone 16B12, Covance) to detect HA-Nse4 in Fig. 2e, 1:2,000 rabbit polyclonal antibodies against Smc5 or Smc6 (ref. 39) (a gift from A. R. Lehmann), 1:1,000 rabbit polyclonal anti-Nse4 (Abgent AP9909A), 1:2,000 rabbit polyclonal anti-Cardif/MAVS (Enzo Life Sciences ALX-210-929), 1:500 mouse monoclonal anti-STAT1 (BD Transduction Laboratories G16920), 1:500 goat polyclonal anti-DDB1 (Everest Biotech), 1:500 rabbit polyclonal anti-HBc (a gift from M. A. Petit), 1:500 mouse monoclonal anti-ubiquitin (Santa Cruz SC-8017), 1:10,000 mouse monoclonal anti-α-tubulin (Sigma-Aldrich T5168), or 1:10,000 mouse monoclonal anti-GAPDH (Sigma-Aldrich G8795) antibodies. Horseradish-peroxidase-conjugated sheep anti-rabbit or anti-mouse, or bovine anti-goat IgG (Amersham Biosciences, 1:5,000), were used as secondary antibodies, and detection was performed with ECL (Pierce).

In Fig. 4b, flash-frozen liver tissue from uPA-SCID mice re-populated with human hepatocytes was homogenized in RIPA buffer containing a broad spectrum protease inhibitor cocktail (Fisher Scientific PI-78430) using a Qiagen Tissue Lyser for mechanical tissue disruption. The resulting homogenates were clarified by centrifugation for 10 min at 4 °C in an Eppendorf 5415D microcentrifuge equipped with a fixed-angle rotor (F45-24-11) at 15,996*g*. Protein concentration was determined using the Bradford Protein assay (Bio-Rad). The membranes were probed with 1:1,000 mouse monoclonal anti-human SMC6 (Abgent AT3956A), 1:1,000 mouse anti-HBsAg (International ImmunoDiagnostics 1113), and 1:1,000 monoclonal mouse anti-human CK-18 (Dako M701029-2). IRDye 680RD goat anti-rabbit or IRDye 800CW goat anti-mouse IgG (Licor, 1:5,000) were used as secondary antibodies. Blots were visualized using an Odyssey Infrared Imaging System (Licor). The species specificity of the anti-human SMC6 and CK-18 antibodies was confirmed by western blot detection of Smc6 and CK-18 in HepG2 cells and PHHs but not in murine hepatocyte AML12 cells (ATCC CRL-2254, data not shown).

In Extended Data Fig. 6g, PHHs were lysed in RIPA buffer supplemented with 1× protease inhibitor cocktail (ThermoScientific), scraped from the plate, and sonicated for 10 s to break up cellular debris. Protein concentration was estimated using a BCA Protein Assay (Novagen). The membranes were probed with 1:1,000 mouse monoclonal anti-Smc5 (Bethyl Laboratories A300-236A), 1:1,000 mouse monoclonal anti-human SMC6 (Abgent AT3956A), 1:1,000 rabbit polyclonal anti-DDB1 (Cell Signaling 5428), and 1:1,000 mouse monoclonal anti-GAPDH (Novus Biologicals G8795). Secondary antibodies and detection were as in Fig. 4b.

RT–PCR analysis. In Extended Data Fig. 2c, total cellular RNA was extracted from HepG2 cells using TRIzol reagent (Invitrogen Life Technologies) following the manufacturer's instructions. RNA was treated with RNase-free DNase (Promega) in the presence of 1 U μl^{-1} RNAsin (Promega). After phenol/chloroform extraction and ethanol precipitation, 1 μg of RNA was reverse transcribed using MLV reverse transcriptase (Promega) and an oligo(dT)15 primer. Quantification by real-time PCR was performed as described[6]. The PCR values were normalized against those obtained for the TBP gene to correct for variation between samples.

In Extended Data Fig. 6e, f, i, total cellular RNA was isolated from PHHs cultured in 96-well plates using an RNeasy 96 Kit (Qiagen) following the manufacturer's instructions. Real-time RT–PCR was performed using a QuantStudio 7 Flex Real-Time PCR System (Invitrogen Life Technologies) following the manufacturer's instructions. The PCR values were normalized against those obtained for the β-actin gene to correct for variation between samples. All oligonucleotide primer sets were manufactured by Life Technologies.

Northern blot analysis. Northern blot analysis in Fig. 3c and Extended Data Fig. 6c was performed on total RNA isolated using TRIzol Reagent (Gibco-Invitrogen) and treated with RNase-Free DNase I (Ambion, Life Technologies). The RNA (10 μg per sample) was denatured by glyoxal treatment and separated on a 1% agarose gel using a NorthernMax-Gly kit (Ambion, Life Technologies). After capillary transfer to Hybond N$^+$ membranes (Amersham), the RNA was fixed on the membrane by ultraviolet cross-linking and hybridized with a ^{32}P-labelled full-length HBV genomic DNA or 28S ribosomal RNA oligonuclotide probes.

HBV virion production. The HBV producing HepG2 cell lines[4] were grown to confluence in DMEM/F12 as described above. When confluence was reached, cells were maintained in a 1:1 mixture of Dulbecco's modified Eagle's medium (DMEM) and Williams' E medium (Invitrogen Life Technologies) supplemented with 5% fetal calf serum, 7×10^{-5} M hydrocortisone hemisuccinate, 5 μg ml^{-1} insulin, and 1% DMSO. HBV-containing supernatants were collected every 2 days for 10 days. The supernatants were pooled, filtered through a 0.22 μm filtration unit (Merck Millipore), and concentrated by overnight precipitation with PEG 8000 (5% final) and centrifugation at 4 °C for 1 h at 6,000*g*. The viral pellet was resuspended in 1/50 of the original culture volume in PBS and sedimented at 4 °C through a 20 ml 10–20% sucrose gradient in PBS at 130,000*g* for 16 h in a SW32 Beckman rotor. The final pellet was resuspended in 1/100 of the original volume in William's E medium supplemented with 2% DMSO and 0.1% SVF and stored in aliquots at −80 °C. Infectious virus titre was estimated by real-time PCR quantification of the viral DNA recovered by immunoprecipitation with an excess of mouse monoclonal antibodies against the large envelope protein (preS1)[40]. Real-time PCR was performed using the SYBR Green PCR Master Mix (Roche Applied Science) and a LightCycler 480 system (Roche Applied Science) as described[34].

HBV virion production from HepAD38 cells was as follows. Cells were grown to confluence in DMEM/F12 supplemented with 10% fetal bovine serum (HyClone, Thermo Scientific), 1% non-essential amino acids, 50 μg ml^{-1} kanamycin, 200 μg ml^{-1} geneticin (all from Gibco Life Technologies), and 0.3 μg ml^{-1} tetracycline (Sigma-Aldrich). Once confluence was reached, the medium was exchanged for identical medium as above, but lacking tetracycline. After 10 days, the virus-containing medium was collected every 3–4 days for 21 days and stored at −80 °C. Subsequently, the aliquots were thawed, precipitated overnight at 4 °C using 6% PEG-8000 (Promega), and centrifuged at 4 °C for 15 min at 1,500*g*. Viral DNA was isolated using a DNeasy Blood & Tissue Kit (Qiagen). Viral titre was determined by real-time PCR.

PHH isolation and HBV infection. PHHs were isolated from resected normal human liver tissue using a two-step perfusion method and cultured as described[4,41]. PHHs were infected with PCR-normalized HBV virus stocks at a multiplicity of infection of 200–400 viral genome equivalents per cell[34]. The experiments in Fig. 3b, e and Extended Data Fig. 6e–i were performed using PHHs purchased from Life Technologies and maintained in William's E medium with added supplements as specified by the vendor. Cells were infected 24 h after plating on collagen-coated 96-well plates (BD Biosciences) with HepAD38-derived HBV virions at 500 viral genome equivalents per cell. In Extended Data Fig. 6g, cells were cultured on collagen-coated six-well plates (BD Biosciences).

Human liver chimaeric uPA-SCID mice. All animal work was performed by Phoenix Bio, in accordance with the Guide for the Care and Use of Laboratory Animals and approved by the Animal Ethics Committee of Phoenix Bio. Nine male uPA-SCID mice were transplanted at 16–23 days old with 1×10^6 PHHs from a single healthy donor as described[42]. After 10–13 weeks, six of these mice were infected by intraperitoneal injection with 5×10^5 genome equivalents of cell-culture-derived HBV genotype C. These animals were killed and serum and liver specimens were collected for measurement of HBV infection 14 weeks later. Serum HBV DNA reached a titre of $>1.5 \times 10^7$ copies per millilitre and serum HBsAg levels were $\geq 3.2 \times 10^2$ IU ml^{-1} in all infected animals. As a control, three of the nine mice were left uninfected. These animals were killed at 15 weeks after transplantation and then processed identically to the HBV-infected animals.

ELISA. The amounts of HBe and HBs antigens present in control and HBV-infected PHH culture media were determined at the indicated times by ELISA using Monolisa HBe Ag-Ab PLUS (Bio-Rad) and Monolisa anti-HBs PLUS (Bio-Rad) kits and a Chemiluminescence Immunoassay (Autobio Diagnostics) kit. In Fig. 3e and Extended Data Fig. 6h, ELISA was performed using an HBeAg EIA

(International Immunodiagnostics) kit and a SpectraMax M5 reader (Molecular Devices). Purified HBeAg was used as standard.

Immunofluorescence and confocal analysis. The proportion of PHHs infected with HBV in Fig. 3d and Extended Data Fig. 5 was estimated by confocal immunofluorescence microscopy for HBsAg expression[34]. PHHs were cultured and infected on collagen-I-coated glass slides. At 9–12 days after infection, cells were fixed with 3% paraformaldehyde in PBS and permeabilized with 0.1% saponin in PBS for 30 min at room temperature. Cells were stained with mouse monoclonal anti-HBs antibodies (a gift from M. A. Petit) diluted 1:200 in PBS containing 0.1% saponin and 3% bovine serum albumin (BSA). After washing in the same buffer, bound antibodies were detected using Alexa-Fluor-647-conjugated goat anti-mouse secondary antibodies (Invitrogen). After additional washes, the slides were mounted with fluorescence mounting medium (ProLong Gold, Life Technologies) and observed under a Leica SP5 X confocal microscope. Image analysis used ImageJ software.

Confocal analysis of Smc6 expression in HBV-infected and uninfected PHHs in Fig. 3b and Extended Data Fig. 7 was performed as follows. PHHs were seeded onto rat tail collagen-I-coated glass coverslips (Corning BioCoat 354089) at a density of 1.2×10^6 cells per well of a 6-well dish and allowed to adhere overnight. Cells were then mock-infected or infected at 500 viral genome equivalents per cell with wild-type HBV derived from HepAD38 cells. On day 13 after infection, the cells were washed twice with PBS and fixed in freshly prepared 4% paraformaldehyde (Electron Microscopy Services RT-15710) in PBS for 10 min at room temperature. After three washes in PBS, the cells were permeabilized in 1% TX-100 in PBS for 10 min at room temperature. Coverslips were incubated overnight at 4 °C in 3% normal goat serum (Jackson ImmunoResearch Laboratories 005-000-121) diluted in PBS to quench non-specific background. Coverslips were then inverted onto 100 μl droplets containing mouse anti-human SMC6 (Abgent AT3956a, 1:500) and polyclonal rabbit anti-HBcAg (Dako B0586, 1:1,600) diluted in 3% normal goat serum, and incubated at room temperature for 1 h. As a control, cells were reacted in parallel with purified mouse IgG2a isotype-matched control (Life Technologies MG2a00). After eight washes in PBS, coverslips were inverted onto 100 μl droplets of highly cross-adsorbed Alexa Fluor 488 goat-anti-mouse (Molecular Probes A11029) or Alexa Fluor 594 goat-anti-rabbit (Molecular Probes A11037) secondary antibodies diluted 1:200 in 3% normal goat serum, and incubated for 1 h at room temperature. Coverslips were washed eight times in PBS and rinsed in double-distilled water before mounting onto glass slides using Vectashield DAPI-containing hardening mounting medium (Vector Laboratories H-1500). Imaging was performed using an upright Zeiss LSM780 confocal system equipped with a ×63 objective lens (NA 1.4) and Zen software. All images within each sample set were captured using identical confocal settings. Images were adjusted for brightness and contrast using Adobe Photoshop CS6.

For immunofluorescence analysis of humanized mouse liver tissues, flash-frozen liver tissues was cryosectioned at a thickness of 5 μm onto microscope slides, fixed for 10 min in ice-cold 4% paraformaldehyde in PBS, and then washed three times in PBS. After one wash in Tween 20-containing AutoWash (BioCare Medical TWB945M), tissue sections were incubated with Rodent Block M solution (BioCare Medical RBM961L) for 30 min at room temperature, then stained overnight at 4 °C in a humidified chamber with mouse monoclonal anti-human Smc6 (Abgent AT3956a, 1:200) and rabbit polyclonal anti-HBsAg (Virostat 1811, 1:500) antibodies diluted in a 1:1 solution of Rodent Block M (BioCare Medical RBM961) and Renoir Red diluent (BioCare Medical PD904M). In parallel, re-population of liver tissue with human hepatocytes was evaluated by staining representative sections from each animal with goat anti-human albumin polyclonal antibodies (Bethyl A80-129A, 1:200), whose species specificity has been demonstrated[43], and for selected animals by co-staining with mouse monoclonal anti-human cytokeratin-18 antibody (Dako M7010, 1:25). Mouse IgG1 (Dako X0931) diluted in Dako Antibody Diluent S0809 served as a negative control (see Extended Data Fig. 9). After three 5 min washes in AutoWash buffer, tissue sections were reacted for 1 h at room temperature with Alexa Fluor 488 donkey anti-mouse and Alexa Fluor 594 donkey anti-rabbit secondary antibodies diluted 1:1,000 in PBS. Coverslips were applied using Vectashield DAPI-containing embedding medium (Vector Laboratories H-1500). Images were acquired using a ×40 objective lens (Fig. 4a) or a ×20 objective lens (Extended Data Fig. 8) and an inverted epifluorescence microscope (Leica DM LB) and arranged using Adobe Photoshop CS6. Note that imaging in Fig. 4a and Extended Data Fig. 8a was limited to areas completely re-populated with human hepatocytes, which were readily distinguished from poorly engrafted areas of mouse liver tissue on the basis of cellular morphology Extended Data Fig. 8b.

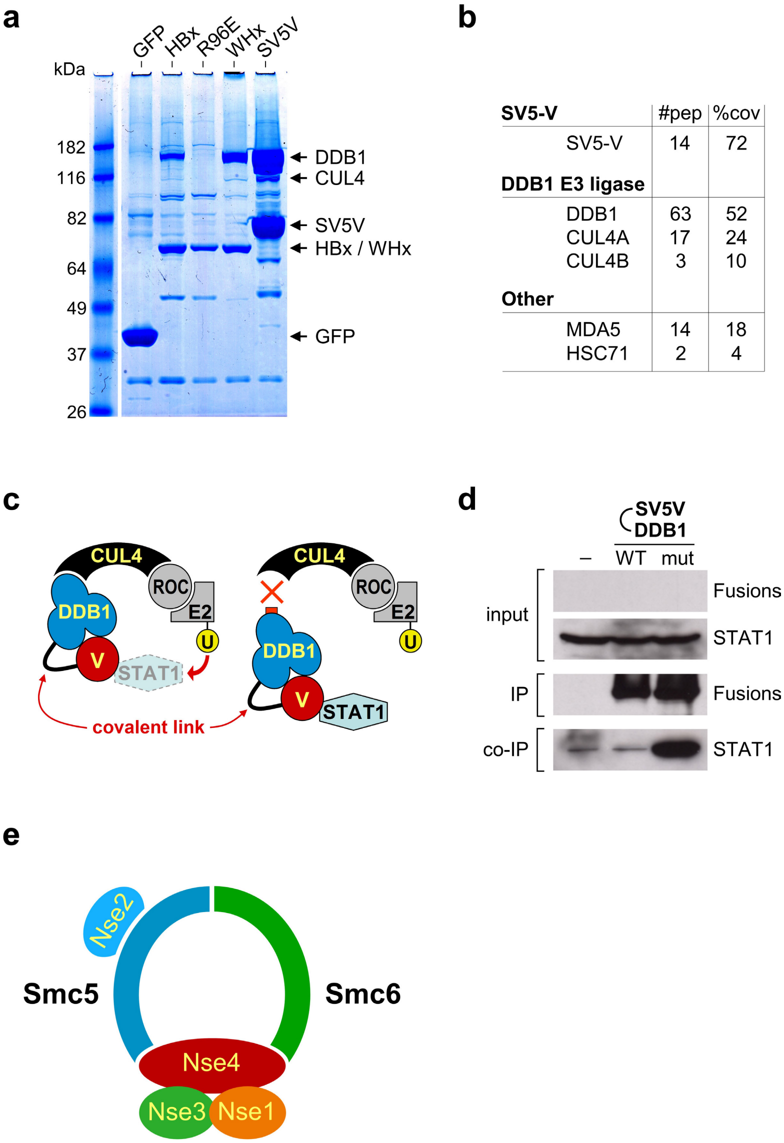
ChIP and quantitative PCR. The ChIP analysis in Fig. 2e was performed using chromatin extracted from about 7×10^6 HepG2 cells cultured in 100 mm diameter wells as above and expressing HA-Nse4 from a lentiviral vector. In Fig. 3f, freshly prepared PHHs were seeded into 75 cm^2 collagen-coated tissue culture flasks (10^7 cells per flask). Two days later, cells were infected with wild-type or HBx-mutant HBV particles. After culture for the indicated periods, cells were fixed with 1% formaldehyde (Sigma Aldrich 47608) for 10 min at 37 °C before quenching with 330 mM glycine. Cells were rinsed twice with ice cold PBS containing EDTA-free protease inhibitors (Roche) and 5 mM Na-butyrate, scraped off in 1.5 ml of the same buffer and collected by centrifugation. Cells were resuspended and lysed for 10 min at 4 °C in 1 ml Nuclear Extraction buffer (10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 1% NP-40) supplemented with protease inhibitor cocktail (Roche). The nuclei were recovered by centrifugation at 500*g* for 5 min at 4 °C in an Eppendorf fixed-angle rotor (FA-45-18-11) and washed once in the same buffer. Nuclei were resuspended in 100 μl FA-lysis buffer (50 mM HEPES/KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100) containing 1% SDS and incubated for 10 min at room temperature. After addition of another 100 μl FA-lysis buffer, the mixture was transferred to 1.5 ml bioruptor microtubes (Diagenode), and sonicated using a Diagenode Bioruptor Pico water bath sonicator (ten cycles of 15 s on and 30 s off at high setting). The sonicated lysates were clarified by centrifugation at 16,000*g* for 10 min and mixed in 800 μl FA-lysis buffer supplemented with protease inhibitors with 40 μl packed bead volume of protein A-Sepharose CL-4B (GE Healthcare) coupled to 4 μl anti-HA antibodies (clone 16B12, Covance) or 6 μl anti-Nse4 antibodies (Abgent AP9909A) and pre-incubated in FA-lysis buffer with sonicated salmon sperm DNA and BSA. After an overnight incubation at 4 °C under constant rotation, the beads were sedimented by brief centrifugation and the supernatant either discarded or in Extended Data Fig. 3c directly mixed and incubated overnight at 4 °C with anti-Nse4 affinity beads as above. The beads were washed twice with 1 ml FA-lysis buffer, twice with FA-lysis buffer adjusted to 500 mM NaCl and 0.5% sodium deoxycholate, and twice with 10 mM Tris-HCl buffer (pH 8.0) containing 1 mM EDTA, 250 mM LiCl, 1% NP-40, and 1% sodium deoxycholate. Bound protein–DNA complexes were released from the HA- and Nse4-affinity resins by incubation for 10 min at 65 °C in 200 μl buffer containing 100 mM Tris-HCl (pH 8.0), 1% SDS, 10 mM EDTA, and 10 mM EGTA. After addition of 200 μl Tris-EDTA (pH 8.0) and 60 μg proteinase K (Bioworld Technology), DNA crosslinks were reversed by overnight incubation at 65 °C. Samples were extracted twice with phenol–chloroform, once with chloroform, and then ethanol precipitated and resuspended in Tris-EDTA buffer. The input DNA treated identically and the recovered DNA were quantified in two separate real-time PCR runs using the KAPA SYBR FAST qPCR Kit Master Mix (2X) Universal (Kapa Biosystems) and the Bio-Rad CFX96 Real-time PCR System. The values were calculated as the ratios between the ChIP signals and the respective input DNA signals. In Fig. 3e, primers specific for the covalently closed circular HBV DNA (cccDNA)[14] were used. Sequences of the oligonucleotide primers are given in Supplementary Table 2.

Statistical analysis. No statistical methods were used to predetermine sample size. Statistical significance was tested using a one-tailed, paired *t*-test (for two sample comparisons) or one-way ANOVA with Dunnett's multiple comparison correction of log-transfomed data (for multiple comparisons). A value of $P < 0.05$ was considered significant. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

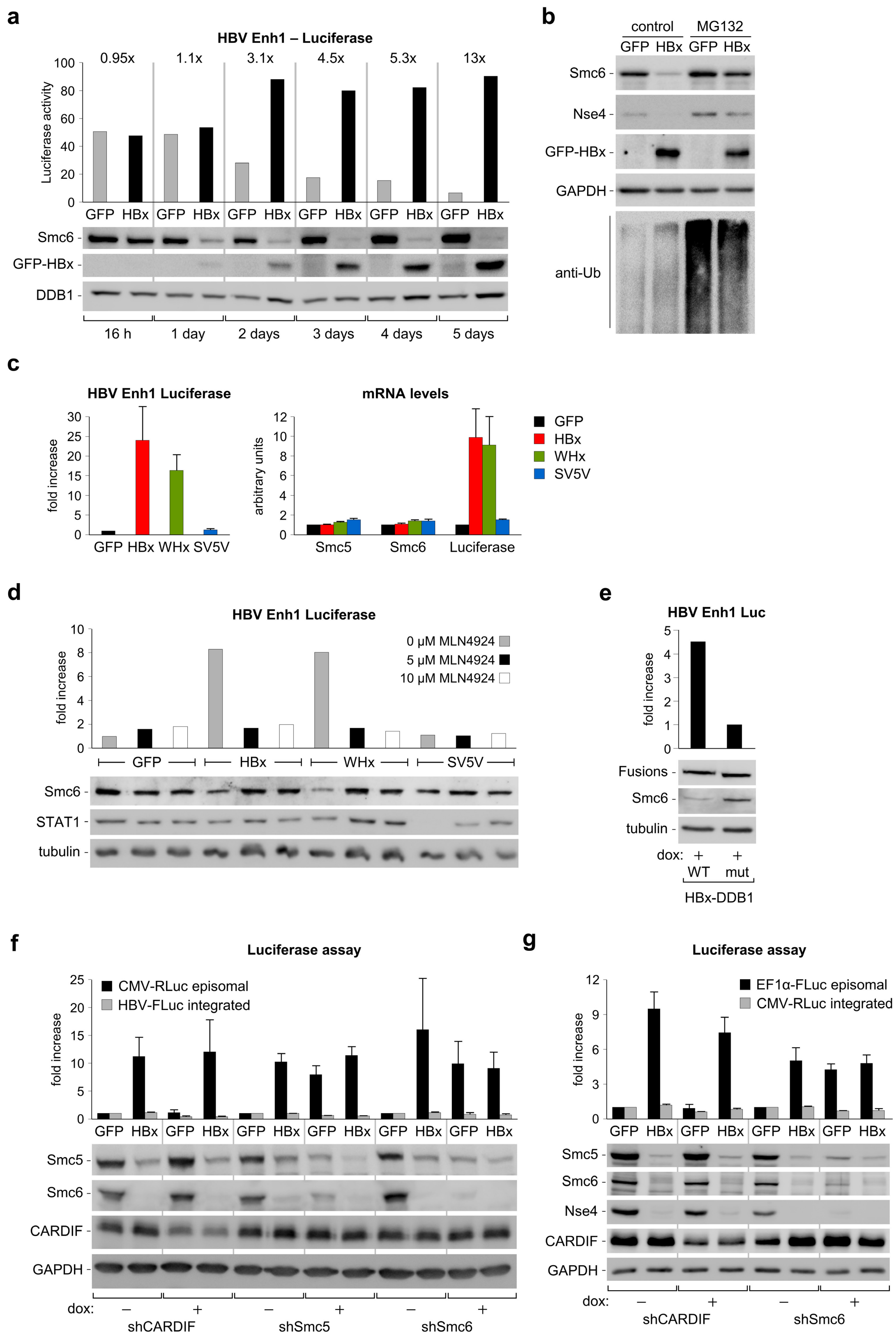
Ethics statement. PHHs were isolated from liver specimens resected from patients undergoing partial hepatectomy (provided by M. Rivoire). Approval from the local and national ethics committees (French Ministry of Research and Education numbers AC-2013-1871 and DC-2013-1870) and informed consent from patients were obtained.

28. Lin-Marq, N., Bontron, S., Leupin, O. & Strubin, M. Hepatitis B virus X protein interferes with cell viability through interaction with the p127-kDa UV-damaged DNA-binding protein. *Virology* **287,** 266–274 (2001).
29. Gloeckner, C. J., Boldt, K., Schumacher, A., Roepman, R. & Ueffing, M. A novel tandem affinity purification strategy for the efficient isolation and characterisation of native protein complexes. *Proteomics* **7,** 4228–4234 (2007).
30. Bontron, S., Lin-Marq, N. & Strubin, M. Hepatitis B virus X protein associated with UV-DDB1 induces cell death in the nucleus and is functionally antagonized by UV-DDB2. *J. Biol. Chem.* **277,** 38847–38854 (2002).
31. Meerbrey, K. L. *et al.* The pINDUCER lentiviral toolkit for inducible RNA interference in vitro and in vivo. *Proc. Natl Acad. Sci. USA* **108,** 3665–3670 (2011).
32. De Iaco, A. & Luban, J. Inhibition of HIV-1 infection by TNPO3 depletion is determined by capsid and detectable after viral cDNA enters the nucleus. *Retrovirology* **8,** 98 (2011).
33. Goldring, C. E. *et al.* Development of a transactivator in hepatoma cells that allows expression of phase I, phase II, and chemical defense genes. *Am. J. Physiol. Cell Physiol.* **290,** C104–C115 (2006).
34. Hantz, O. *et al.* Persistence of the hepatitis B virus covalently closed circular DNA in HepaRG human hepatocyte-like cells. *J. Gen. Virol.* **90,** 127–135 (2009).

35. Ladner, S. K. *et al.* Inducible expression of human hepatitis B virus (HBV) in stably transfected hepatoblastoma cells: a novel system for screening potential inhibitors of HBV replication. *Antimicrob. Agents Chemother.* **41,** 1715–1720 (1997).
36. Kutner, R. H., Zhang, X. Y. & Reiser, J. Production, concentration and titration of pseudotyped HIV-1-based lentiviral vectors. *Nature Protocols* **4,** 495–505 (2009).
37. Rothe, M. *et al.* Epidermal growth factor improves lentivirus vector gene transfer into primary mouse hepatocytes. *Gene Ther.* **19,** 425–434 (2012).
38. Martin-Lluesma, S. *et al.* Hepatitis B virus X protein affects S phase progression leading to chromosome segregation defects by binding to damaged DNA binding protein 1. *Hepatology* **48,** 1467–1476 (2008).
39. Taylor, E. M. *et al.* Characterization of a novel human SMC heterodimer homologous to the *Schizosaccharomyces pombe* Rad18/Spr18 complex. *Mol. Biol. Cell* **12,** 1583–1594 (2001).
40. Chemin, I. *et al.* Correlation between HBV DNA detection by polymerase chain reaction and Pre-S1 antigenemia in symptomatic and asymptomatic hepatitis B virus infections. *J. Med. Virol.* **33,** 51–57 (1991).
41. Gripon, P., Diot, C. & Guguen-Guillouzo, C. Reproducible high level infection of cultured adult human hepatocytes by hepatitis B virus: effect of polyethylene glycol on adsorption and penetration. *Virology* **192,** 534–540 (1993).
42. Fujiwara, S. *et al.* A novel animal model for *in vivo* study of liver cancer metastasis. *World J. Gastroenterol.* **18,** 3875–3882 (2012).
43. Ishida, Y. *et al.* Novel robust *in vitro* hepatitis B virus infection model using fresh human hepatocytes isolated from humanized mice. *Am. J. Pathol.* **185,** 1275–1285 (2015).
44. Andrejeva, J. *et al.* The V proteins of paramyxoviruses bind the IFN-inducible RNA helicase, mda-5, and inhibit its activation of the IFN-beta promoter. *Proc. Natl Acad. Sci. USA* **101,** 17264–17269 (2004).

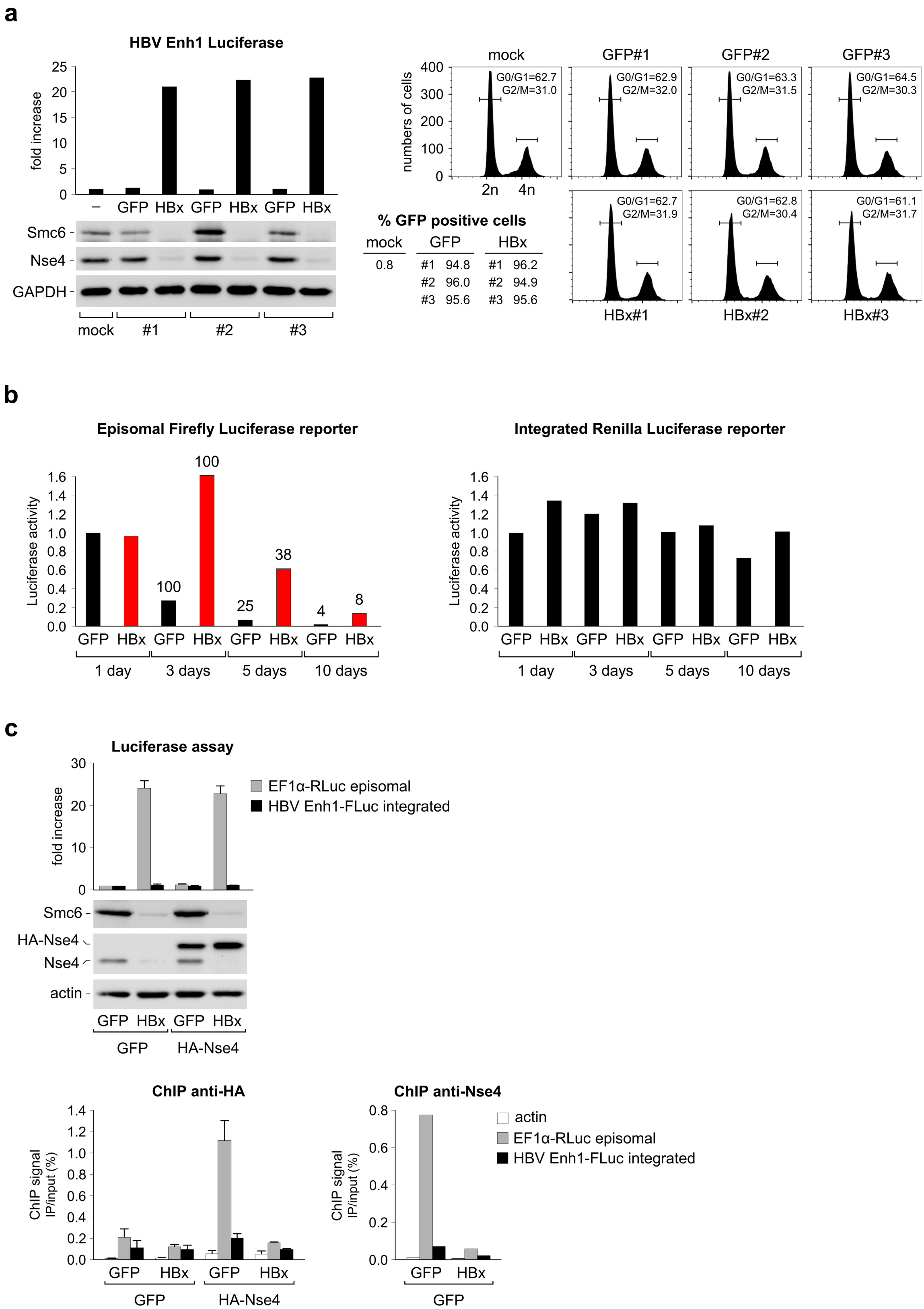


Extended Data Figure 1 | A strategy to identify E3 ubiquitin ligase substrates. a, Stable HepG2 cell lines were generated expressing GFP or the indicated GFP-tagged viral proteins carrying an N-terminal tandem Flag/StrepII (FS) tag from a doxycycline-inducible promoter (ref. 29 and see Methods). After doxycycline induction, whole-cell extracts were prepared and the FS-tagged and associated proteins were purified by TAP. Eluted proteins were separated by gel electrophoresis and stained with Coomassie blue. CUL4 was identified by mass spectrometry. No protein candidate was identified showing a profile predicted for an HBx substrate: that is, co-purifying specifically with HBx and WHx but not with SV5-V, or co-purifying only with the DDB1-binding-defective HBx(R96E) mutant because of its otherwise unstable interaction with HBx and WHx when recruited to the E3 ligase. In particular, the Smc5/6 complex was not detected in these purifications. **b**, Purification of SV5-V and associated proteins. FS-tagged SV5-V was purified from a stably expressing HepG2 cell line by TAP as above. Proteins were eluted under native conditions and processed for analysis by peptide fragmentation sequencing (nano-LC–ESI MS/MS). The number of unique peptides for each protein and the percentage of total protein sequence covered by these peptides are indicated. Listed are proteins identified with 100% certainty and represented by at least two peptides identified at a 95% confidence level. CUL4A and CUL4B are closely related paralogs. Note that no peptides were detected, even at low confidence level, for STAT1 that is well known to be recruited by SV5-V to the DDB1 E3 ligase for ubiquitin-mediated degradation. By contrast MDA5, a cytoplasmic sensor of viral RNA to which SV5-V also binds but, in contrast to STAT1, without causing its degradation[44], was present in the SV5-V purifications. This suggests that, in contrast to MDA5, STAT1 associates with SV5-V very transiently, presumably because of its rapid degradation or dissociation from the E3 ligase complex. We considered that the same may be true for the HBx target, thus precluding its identification by regular affinity purification. **c**, As a proof of principle for the use of a fusion strategy to identify the HBx target, we covalently linked SV5-V to wild-type DDB1 or to a CUL4-binding-defective DDB1 mutant that could not incorporate in the E3 ligase complex. See text for details. **d**, HepG2 cells were mock transfected (−) or transfected with plasmid DNA expressing the indicated FS-tagged SV5-V–DDB1 fusions. Whole-cell extracts were prepared and the fusion proteins purified by a single round of affinity purification. The amounts of fusion proteins recovered (IP) and the presence of STAT1 in the eluates (co-IP) were assessed by western blotting. The fusion proteins were revealed using anti-Flag antibodies and were expressed at too low levels for detection in the crude extracts (input). Note that high amounts of STAT1 are recovered only with the mutant SV5-V–DDB1 fusion. **e**, Architecture of the Smc5/6 complex. The core of the Smc5/6 complex is formed by a heterodimer of Smc5 and Smc6. The two proteins form a V-shaped structure and associate with four non-SMC proteins, designated Nse1–Nse4 (refs 15, 17). Note that depletion of any of the Smc5/6 complex subunit, other than Nse2, results in destabilization and degradation of the entire complex[21].



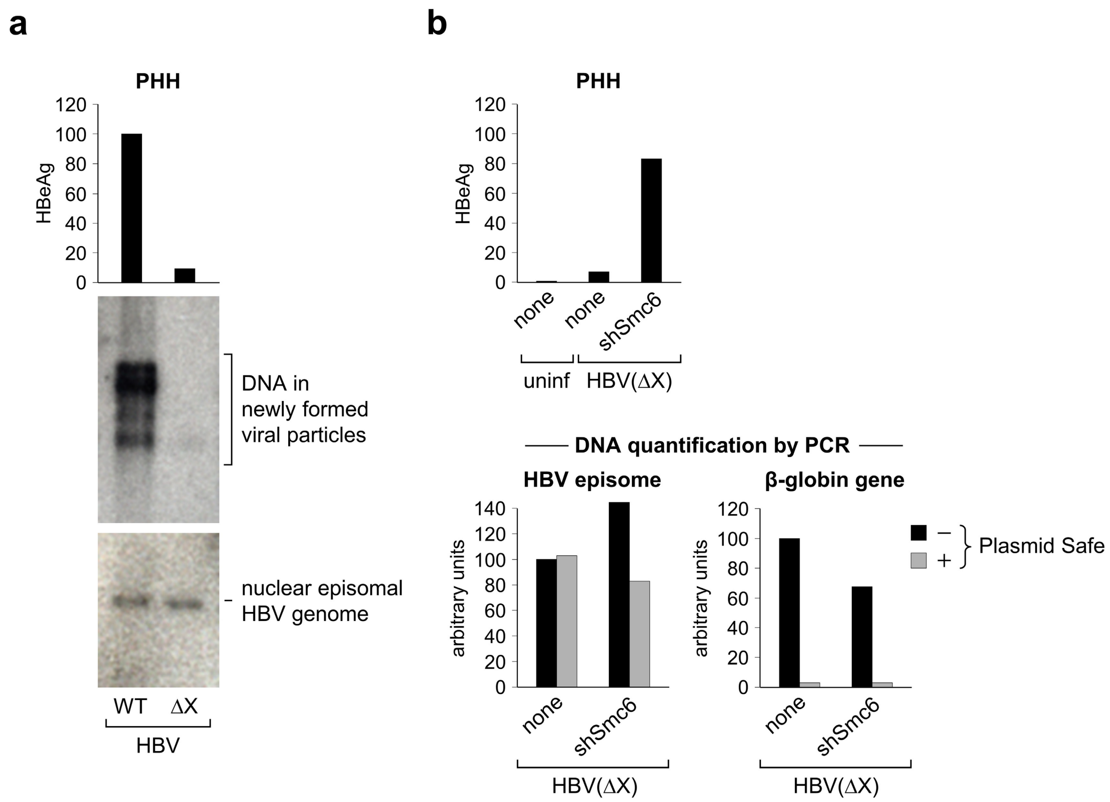
Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | HBx induces rapid depletion of the Smc5/6 complex by an MG132 proteasome inhibitor-sensitive pathway to stimulate extrachromosomal reporter gene activity. **a**, Same experiment as in Fig. 2a except that luciferase activity and Smc6 protein levels were monitored at the indicated time points after transduction of GFP or GFP–HBx. Luciferase activities are expressed in arbitrary units, with the fold increase in HBx-expressing cells relative to GFP control cells indicated on top. GFP–HBx was detected using anti-GFP antibodies. One out of two independent experiments. **b**, HepG2 cells were transduced with lentiviral vectors encoding GFP or GFP–HBx. Then, 16 h later, DMSO (control) or 10 μM MG132 was added to the culture to inhibit proteasome activity. Cells were harvested 8 h later and the level of the indicated proteins and of global ubiquitination was analysed by western blotting. GFP–HBx was detected using anti-GFP antibodies. One out of two independent experiments. **c**, The effect of the indicated viral proteins on Smc5, Smc6, and luciferase mRNA levels was determined by real-time RT–PCR. The values are relative to those measured in GFP-transduced cells, which were set to 1. **d**, Similar experiment as in Fig. 2b except that two concentrations of MLN4924 were used. **e**, Similar experiment as in Fig. 2c. **f**, Similar experiment as in Fig. 2d but with HepG2 cells containing the HBV Enhancer-I-driven firefly luciferase reporter gene (HBV-FLuc) integrated into the chromosome and an episomal *Renilla* luciferase construct driven by the CMV promoter (CMV-RLuc). Note that the shRNAs against Smc5 and Smc6 used in this experiment differ from those in Fig. 2d (see Methods). **g**, Same with the *Renilla* luciferase CMV promoter construct chromosomally integrated and a firefly luciferase reporter gene driven by the EF1α promoter episomal. Data in **c**, **f** and **g** represent the mean ± s.e.m. of three independent experiments.

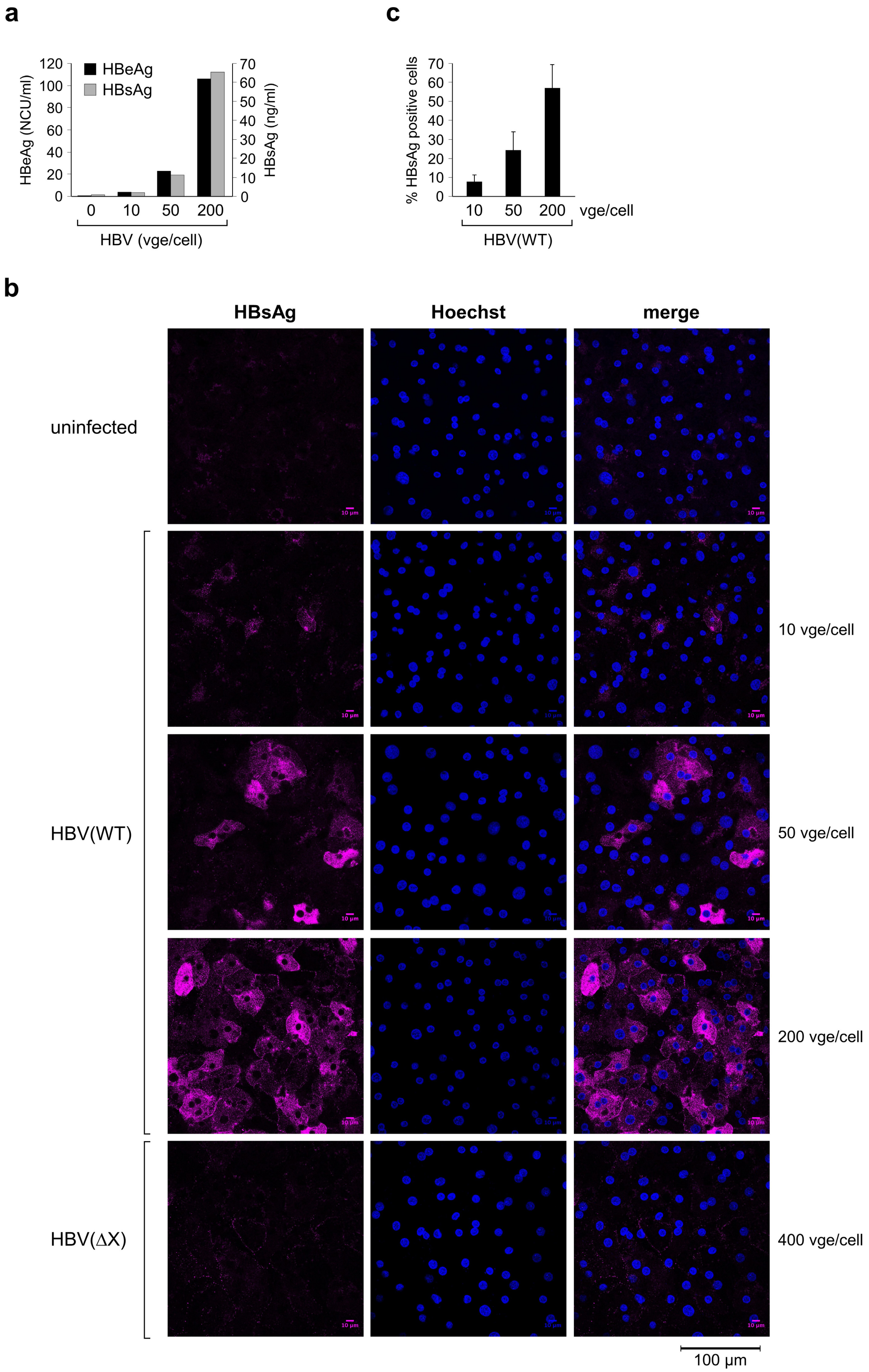


Extended Data Figure 3 | See next page for caption.

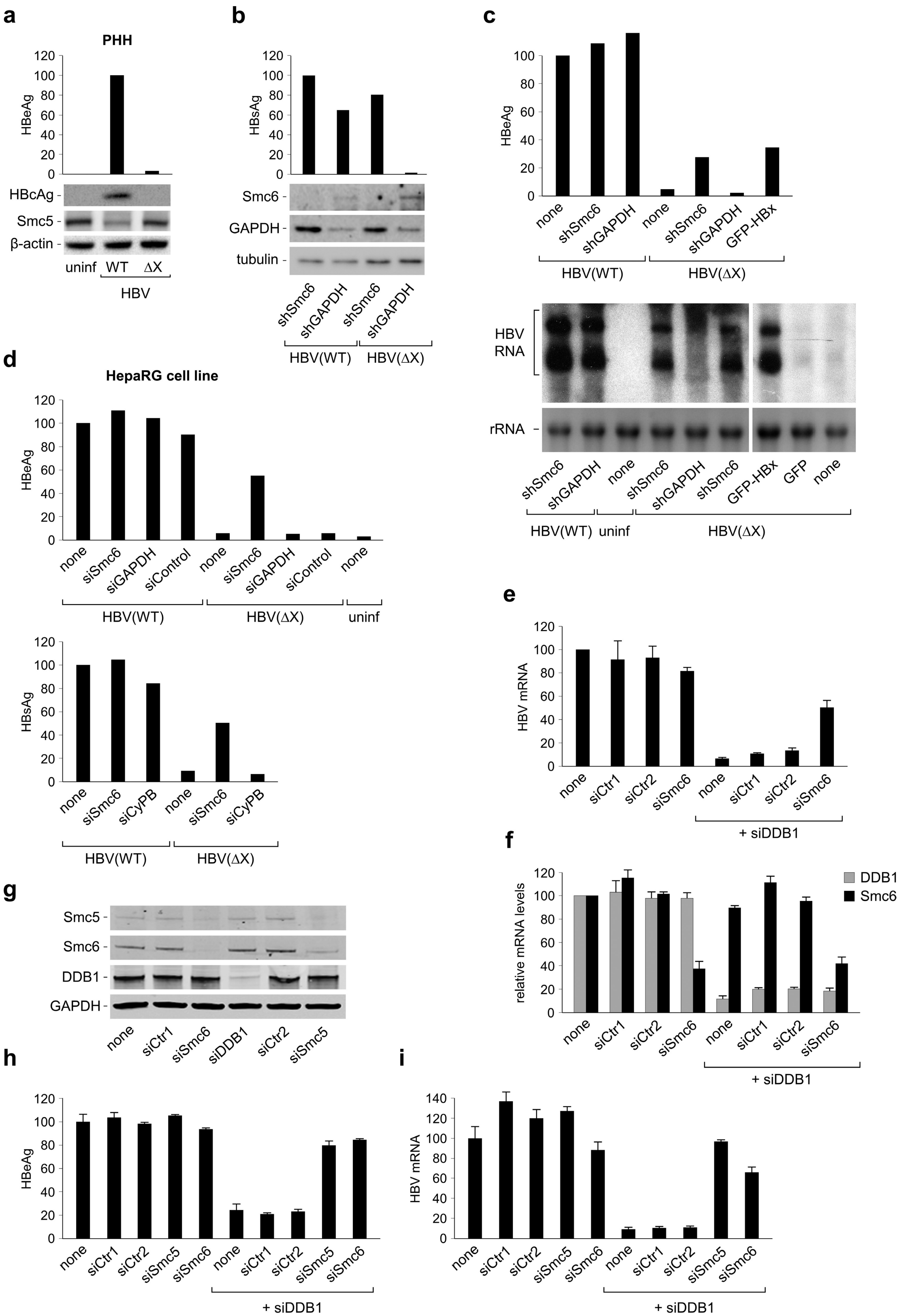
Extended Data Figure 3 | Smc5/6 degradation by HBx does not alter the cell cycle or promote chromosomal integration of the reporter but prevents the binding of Smc5/6 to the episomal DNA template. **a**, HepG2 cells were transfected with the HBV Enh1 luciferase reporter plasmid, split, and then either mock-transduced or transduced in triplicates with lentiviral vectors encoding GFP or GFP–HBx. Luciferase activity and protein expression (left panel) were monitored as in Fig. 2 at 5 days after transfection. In parallel, cells were analysed by FACS for GFP positivity (middle inset) and for DNA content (right panels) after propidium iodide staining. **b**, HBx does not promote stable integration of the reporter gene into chromosomes. HepG2 cells were transduced with a lentiviral vector to chromosomally integrate a *Renilla* luciferase reporter gene and subsequently transfected with a firefly reporter construct. Cells were then split equally and transduced with lentiviral vectors encoding GFP or GFP–HBx. Activity of the episomal firefly (left) and integrated *Renilla* (right) luciferase genes was monitored at the indicated time points after transduction of GFP or GFP–HBx. The values are expressed in arbitrary units relative to those measured at day 1 in the GFP control, which was set to 1.0. Indicated above the columns in the left panel is the remaining luciferase activity (expressed in per cent) relative to that measured 3 days after transfection. Note that in the absence of HBx (black bars in the left panel) the episomal firefly luciferase signal decreases rapidly with time, presumably because of loss of the reporter plasmid. In the presence of HBx (red bars in the left panel), the signal slightly increases to reach sixfold higher levels relative to the GFP control at 3 days after transfection, and then drops with kinetics close to that observed in the GFP control. In contrast, expression of the integrated *Renilla* reporter remains constant (right panel). This argues against HBx stimulating episomal reporter activity by promoting its integration into chromosomes. **c**, Left panels, the same ChIP experiment using anti-HA antibodies as in Fig. 2e but with HepG2 cells containing the HBV Enhancer-I-driven firefly luciferase reporter (HBV Enh1-FLuc) integrated into the chromosome and an episomal EF1α *Renilla* luciferase construct (EF1α-RLuc) as a control. Data represent the mean $\pm$ s.e.m. of three independent experiments. In one experiment, the unbound material from the indicated samples expressing no HA-tagged Nse4 was further purified using anti-Nse4 antibodies (lower right). Data are expressed as the percentage of input DNA recovered by ChIP. Note that in this experiment, the episomal reporter was delivered using an integrase-defective lentiviral vector[32].



Extended Data Figure 4 | HBx or Smc5/6 knockdown does not affect HBV genome copy number. a, PHHs were infected with normalized stocks of wild-type (WT) or HBx-deficient (ΔX) HBV as in Fig. 3. HBeAg secretion (top) was assessed 10 days later by ELISA. In parallel, the amounts of cytoplasmic DNA replicative intermediates produced during reverse transcription of the viral pregenomic RNA (middle) and of the nuclear episomal HBV template extracted using a modified Hirt procedure (bottom) were analysed by Southern blot as reported[34]. **b**, PHHs were either mock transduced (none) or transduced with a lentiviral construct expressing an shRNA specific for Smc6 and the next day infected with HBx-deficient HBV particles (HBV(ΔX) as in Fig. 3c. HBeAg secretion was assessed 12 days later by ELISA (top). In parallel, the amount of nuclear HBV genome was quantified by real-time FRET–PCR[34] both directly or after treatment with Plasmid-safe DNase (Epicentre), an exonuclease that degrades single-stranded and double-stranded linear DNA but not the episomal HBV DNA (bottom left). Shown as a control for Plasmid-safe DNase treatment are the results for the chromosomal β-globin gene (bottom right).

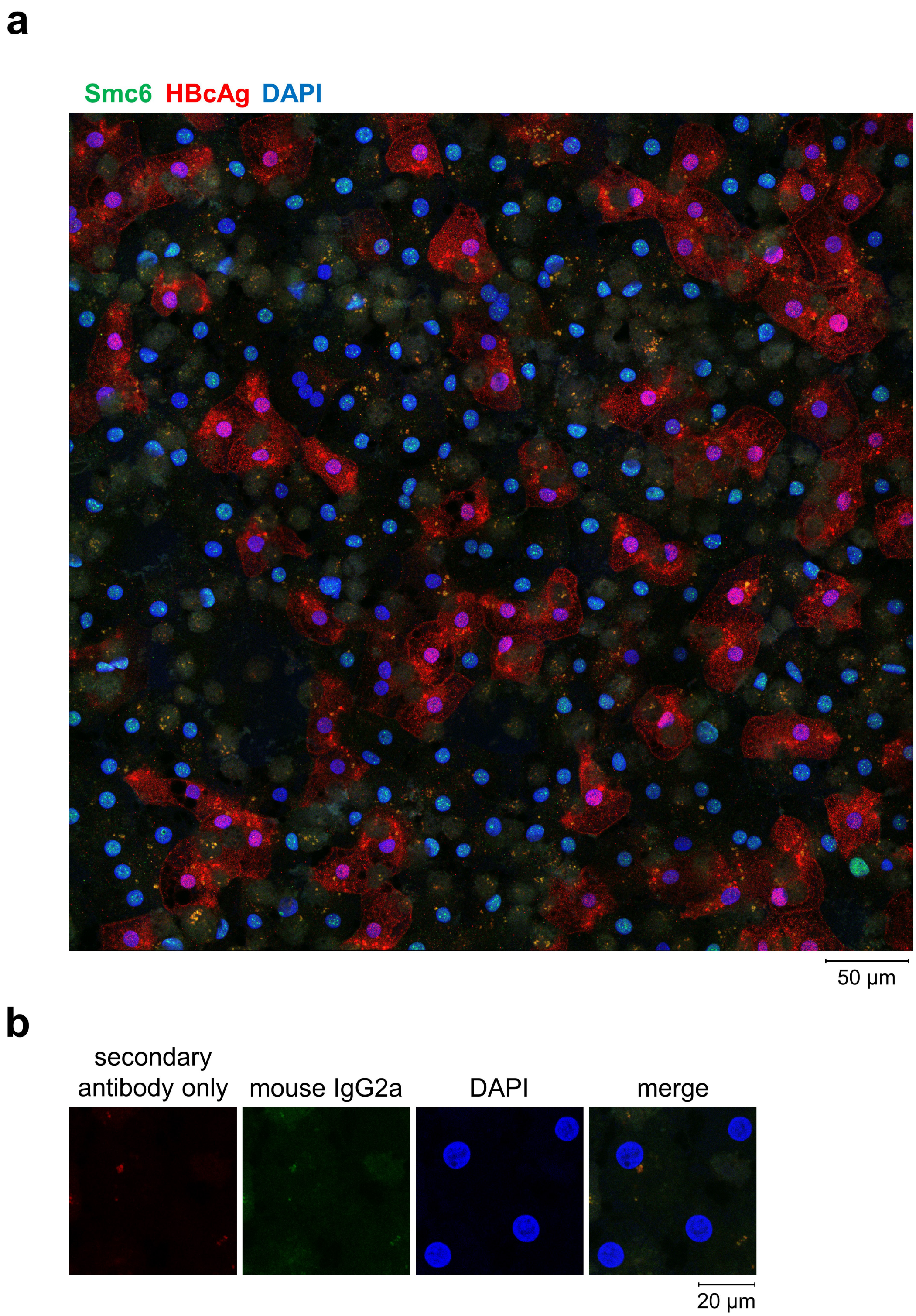


Extended Data Figure 5 | Efficiency of PHH infection by HBV. **a**, Purified PHHs were either left uninfected (0) or infected with wild-type HBV particles at the indicated viral-genome equivalents per cell (see Methods). HBeAg and surface antigen (HBsAg) secretion into the culture supernatants was assessed 10 days later by ELISA. Concentrations are expressed in, respectively, national clinical units (NCU) per millilitre (HBeAg) and nanograms per millilitre (HBsAg) and were determined according to the manufacturers' guidelines (Autobio Diagnostics). The noise signals measured with the supernatants from the uninfected cells were 0.50 NCU ml^{-1} and 0.88 ng ml^{-1}. **b**, PHHs infected with normalized stocks of wild-type (WT) or HBx-deficient (ΔX) HBV at the indicated viral genome equivalents per cell were examined for HBsAg expression at 9 days after infection by indirect immunofluorescence confocal microscopy. Cell nuclei were stained with Hoechst dye (blue). **c**, Quantification of confocal images. Data are mean ± s.d. of at least three fields. Note that infection in Fig. 3 was with 200 viral genome equivalents per cell or more.

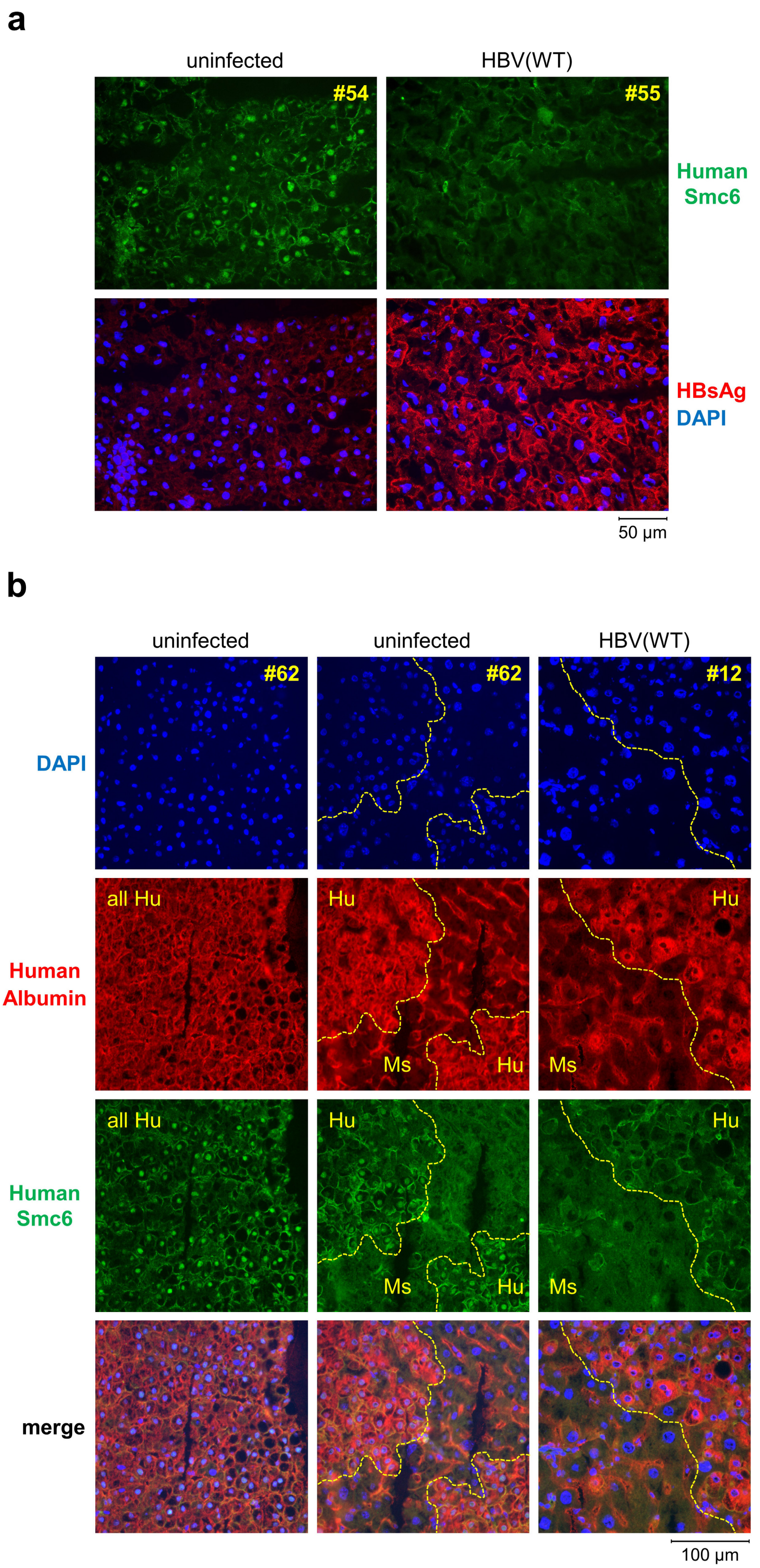


Extended Data Figure 6 | See next page for caption.

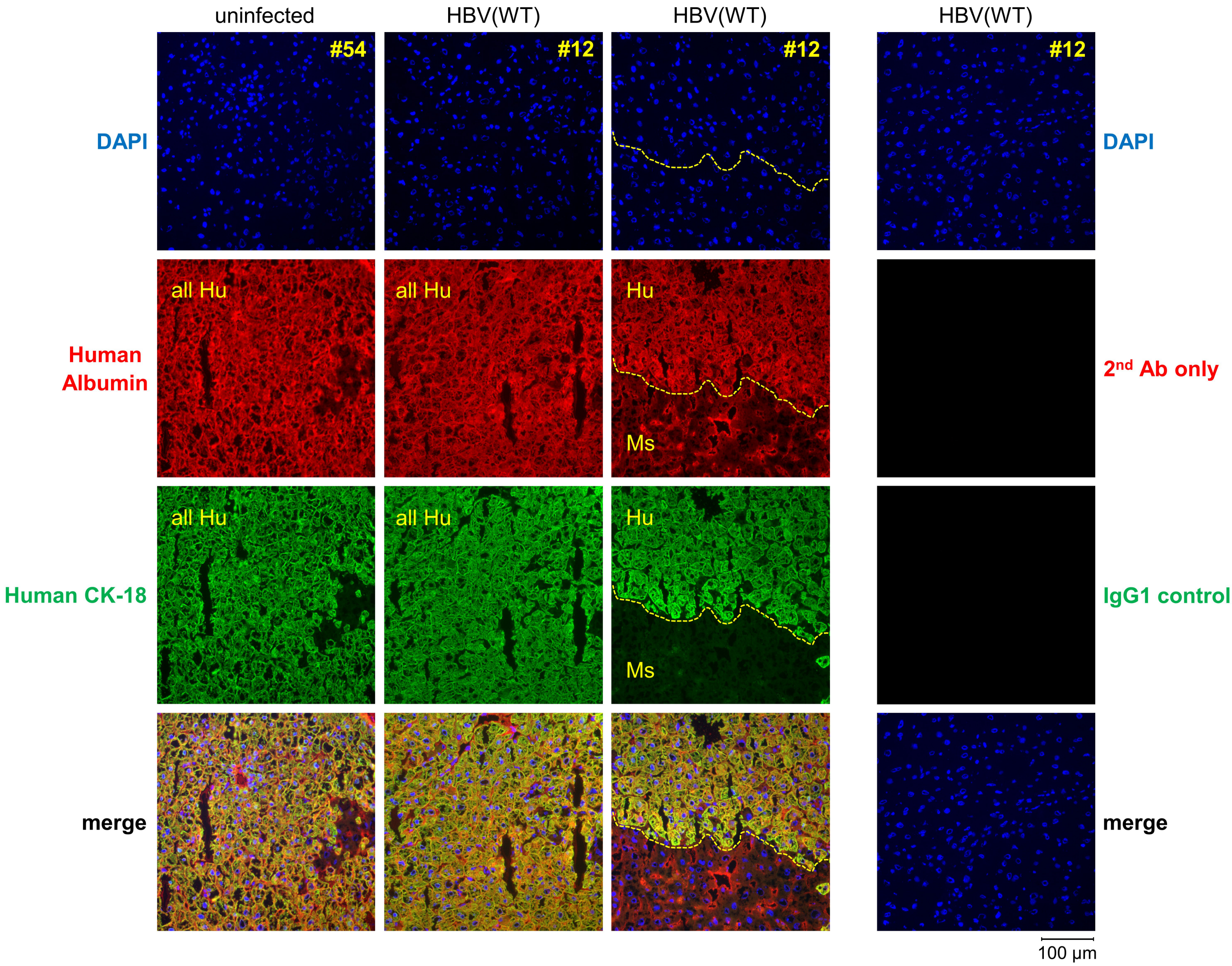
Extended Data Figure 6 | Silencing of Smc5/6 restores HBx-negative HBV transcription and rescues wild-type HBV on a DDB1 knockdown background. a, Biological replicate of Fig. 3a using PHHs from a different donor. **b**, Control for lentiviral shRNA-mediated depletion in PHHs. PHHs were transduced with lentiviral constructs expressing the indicated shRNAs and infected with HBV as in Fig. 3c. HBsAg secretion and the amounts of the indicated proteins were assessed 9 days later by ELISA and western blot analysis. HBsAg concentrations are relative to those measured in mock-transduced cells infected with wild-type HBV, which were taken as 100 (not shown). **c**, Independent northern blot analysis of HBV RNA production as in Fig. 3c using PHHs from a different donor. All lanes are from the same gel and exposure. Shown in the upper panel is the corresponding ELISA for HBeAg secretion. **d**, Smc5/6 silencing restores expression of HBx-negative HBV in HepaRG cells. Differentiated HepaRG cells were either left uninfected or infected with normalized stocks of wild-type (WT) or HBx-deficient (ΔX) HBV particles. Twenty-four hours later, cells were either mock transfected (none) or transfected with the indicated siRNA. HBeAg secretion was measured 10 days after infection by ELISA. The values are in arbitrary units relative to the wild-type HBV control, which was given a value of 100. Shown in the upper and lower panels are two independent experiments. CyPB, cyclophilin B. **e**, HBV mRNA expression. Total RNA was extracted from the samples analysed for HBeAg secretion in Fig. 3e. HBV mRNA levels were measured by real-time RT–PCR. The values normalized to β-actin are given in arbitrary units relative to those measured in mock-transfected cells, which were set to 100. Data represent the mean ± s.e.m. of $n = 4$ independent experiments performed with two different PHH donors. **f**, Control of siRNA efficacy. DDB1 (grey bars) and Smc6 (black bars) mRNA levels in Fig. 3e were measured by real-time RT–PCR and normalized to β-actin as in **e**. The values are given in arbitrary units relative to those measured in mock-transfected cells, which were set to 100. Data represent the mean ± s.e.m. of $n = 4$ independent experiments performed with two different PHH donors. **g**, Control western blot analysis of siRNA-mediated protein depletion. PHHs were transfected with the indicated siRNAs 3 days after plating. Protein levels were measured 10 days later. **h**, Data of one of the four experiments used in Fig. 3e that includes an siRNA targeting Smc5. Mean ± s.e.m. of triplicate measurements. **i**, Same experiment as in **h** but measuring HBV mRNA levels. Mean ± s.e.m. of triplicate measurements.



Extended Data Figure 7 | HBV infection induces Smc6 degradation in PHHs. a, Same as in Fig. 3b but 4 × 4 contiguous images were acquired and stitched together to produce a single image for examination of 200–300 PHHs. Shown are Smc6 (green), HBcAg (red), and DAPI (blue). **b**, Representative images of uninfected PHHs stained with mouse isotype-control-matched IgG2a primary antibodies or Alexa Fluor 488- and 594-conjugated secondary antibodies alone. Samples were imaged and processed as in Fig. 3b.



Extended Data Figure 8 | HBV infection induces Smc6 degradation in humanized mouse liver tissue. **a**, Same as in Fig. 4a but with different animal samples. **b**, Specificity of detection of human Smc6 in humanized mouse liver tissues. Fresh-frozen uninfected (animal 62) and infected (animal 12) humanized mouse liver tissues were stained with DAPI nuclear stain (blue) and with antibodies against human albumin (red) and human Smc6 (green). In the middle and right panels, a yellow dotted line delineates the interface between the human (Hu) and mouse (Ms) hepatocyte populations.



Extended Data Figure 9 | Human albumin and human cytokeratin-18 as markers for human hepatocytes in humanized mouse liver tissues. Fresh-frozen uninfected (animal 54) or infected (animal 12) humanized mouse liver tissues were stained with DAPI nuclear stain (blue) and with antibodies against human albumin (red) and human cytokeratin-18 (CK-18, green). A yellow dotted line delineates the interface between the human (Hu) and mouse (Ms) hepatocyte populations in the third column from left. Shown on the right are control staining of liver tissue from animal 12 with secondary antibody only or IgG1 isotype control.

doi:10.1038/nature17150

Sequence-dependent but not sequence-specific piRNA adhesion traps mRNAs to the germ plasm

Anastassios Vourekas[1]*, Panagiotis Alexiou[1]*, Nicholas Vrettos[1], Manolis Maragkakis[1] & Zissimos Mourelatos[1]

The conserved Piwi family of proteins and piwi-interacting RNAs (piRNAs) have a central role in genomic stability, which is inextricably linked to germ-cell formation, by forming Piwi ribonucleoproteins (piRNPs) that silence transposable elements[1]. In *Drosophila melanogaster* and other animals, primordial germ-cell specification in the developing embryo is driven by maternal messenger RNAs and proteins that assemble into specialized messenger ribonucleoproteins (mRNPs) localized in the germ (pole) plasm at the posterior of the oocyte[2,3]. Maternal piRNPs, especially those loaded on the Piwi protein Aubergine (Aub), are transmitted to the germ plasm to initiate transposon silencing in the offspring germ line[4–7]. The transport of mRNAs to the oocyte by midoogenesis is an active, microtubule-dependent process[8]; mRNAs necessary for primordial germ-cell formation are enriched in the germ plasm at late oogenesis via a diffusion and entrapment mechanism, the molecular identity of which remains unknown[8,9]. Aub is a central component of germ granule RNPs, which house mRNAs in the germ plasm[10–12], and interactions between Aub and Tudor are essential for the formation of germ granules[13–16]. Here we show that Aub-loaded piRNAs use partial base-pairing characteristics of Argonaute RNPs to bind mRNAs randomly in *Drosophila*, acting as an adhesive trap that captures mRNAs in the germ plasm, in a Tudor-dependent manner. Notably, germ plasm mRNAs in drosophilids are generally longer and more abundant than other mRNAs, suggesting that they provide more target sites for piRNAs to promote their preferential tethering in germ granules. Thus, complexes containing Tudor, Aub piRNPs and mRNAs couple piRNA inheritance with germline specification. Our findings reveal an unexpected function for piRNP complexes in mRNA trapping that may be generally relevant to the function of animal germ granules.

We performed ultraviolet crosslinking followed by stringent immunoprecipitation (CLIP)[17] for Aub (Fig. 1a) and standard small RNA immunoprecipitation, using a highly specific antibody that we generated (Extended Data Fig. 1a) from wild-type (*yw*) ovaries and from *yw* and Tudor-null (*tud*) *Drosophila* embryos collected up to 2 h after laying (0–2-h embryos); this is before zygotic transcription and degradation of maternal mRNAs. Crosslinked RNA–Aub complexes yielded strong, specific signals that were absent from non-immune serum and no-ultraviolet controls (Fig. 1a). CLIP and immunoprecipitation libraries contained essentially identical 23–29-nucleotide piRNAs (Fig. 1b, Extended Data Figs 1b–g and 2a–f and Extended Data Table 1). We verified minimal changes in the piRNA load of Aub in *tud* versus *yw* ovaries[13] (Extended Data Fig. 2g), and found no changes in the piRNA load of 0–2-h embryos compared to ovaries in both genotypes (Extended Data Fig. 2h, i). Larger CLIP tags (lgCLIPs, ≥36 nucleotides) are present in libraries prepared from larger RNP complexes (Fig. 1a–c, Extended Data Fig. 1d and Supplementary Results).

We observe considerable overlap of retrotransposon lgCLIPs with complementary piRNAs (Extended Data Fig. 3a and Supplementary Table 1) and strong positive correlation of their abundances (Extended Data Fig. 3b, c). Relative distance analysis reveals high occurrence of lgCLIPs with a 10-nucleotide overlap to complementary piRNAs (Fig. 1d, peak at position +9) for all three genotypes. The majority of such lgCLIPs bear an adenine at the tenth position (Fig. 1e), and show prominent 5′–5′ end coincidence with Ago3 piRNAs (Fig. 1f), indicating that they correspond to ping-pong intermediate fragments produced by Aub slicing[1]. Furthermore, a second peak at position −15 (Fig. 1d), which is 25 nucleotides (the median Aub piRNA length) from position +9, represents 5′ ends of fragments of trigger piRNA targets undergoing phased piRNA biogenesis[18]. The above results indicate that CLIP captures piRNA biogenesis, complementary retrotransposon targeting and the transient products of Aub slicing activity (Fig. 1g).

A large percentage (∼50–66%) of lgCLIPs from all CLIP libraries are mRNA-derived (Fig. 1c and Extended Data Fig. 1g). Most Aub-bound mRNAs are not substrates for piRNA processing (Extended Data Fig. 4a). The Aub lgCLIP density is relatively high within 3′ untranslated regions (UTRs) compared to RNA sequencing (RNA-seq) analysis, and overall lgCLIP abundance is not correlated with mRNA abundance (Extended Data Fig. 4b–d), suggesting specific target mRNA recognition. We cross-indexed Aub-bound mRNAs with the mRNA localization categories (compiled in ref. 19). Notably, posterior localization categories are significantly enriched in all three sets of Aub CLIP libraries (embryo: *yw* and *tud*, ovary: *yw*) (Supplementary Table 2). Most importantly, we find 15 posterior and germ-cell localization categories significantly depleted, and ubiquitous mRNAs enriched in *tud* embryo compared to *yw* embryo CLIP libraries (Supplementary Table 3). Posteriorly localized mRNAs appear marginally upregulated compared to other localization categories in *tud* versus *yw* embryo RNA-seq libraries (two-sided *t*-test, $P = 0.01594$), ruling out the possibility that the reduced Aub binding is due to reduced posterior mRNA levels in *tud* embryos. Both Aub (Extended Data Fig. 1a) and germ plasm mRNAs[15,20] are uniformly distributed throughout *tud* embryos; therefore, the observed loss of binding specificity towards posterior mRNAs in the absence of Tudor can only be attributed to the disruption of the germ plasm. Thus, our experimental approach allows the identification of the mRNAs specifically bound by Aub in the germ plasm, irrespective of the function of Aub in the clearance of maternal mRNAs in the somatic part of the embryo[21,22]. To identify the primary mRNA targets of Aub within the germ plasm during the formation of germ cells, we calculated the rank product of the normalized lgCLIP values for mRNAs in the 12 posterior localization categories marked with an asterisk in Supplementary Table 3, from three replicate *yw* embryo libraries ($P < 0.05$). The list contains 220 genes, many of which appear enriched or selectively protected in germ cells[10], and with established roles in germ-cell specification and development such as *cycB*, *nos*, *osk*, *gcl*, *pgc* and *Hsp83* (Supplementary Table 4). Characterization of Aub RNPs from early embryos provides independent support for the association of germ plasm mRNAs with Aub (Supplementary Results

[1]Department of Pathology and Laboratory Medicine, Division of Neuropathology, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine; PENN Genome Frontiers Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
*These authors contributed equally to this work.

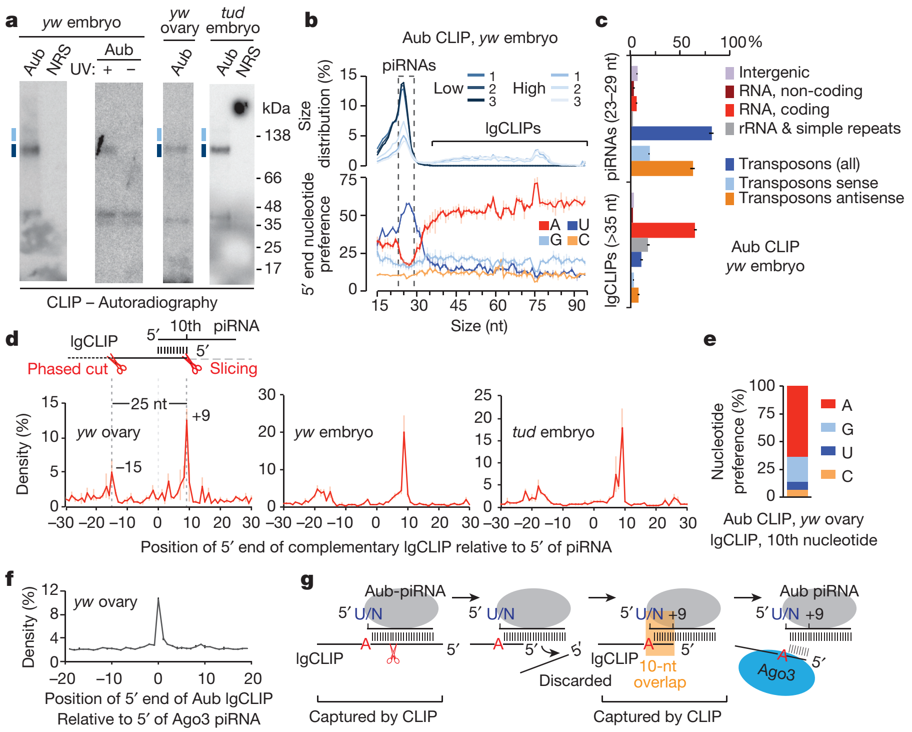


Figure 1 | Transcriptome-wide identification of RNAs bound by Aub and *in vivo* retrotransposon targeting and slicing captured by CLIP. **a**, Aub CLIPs; separate libraries were prepared from RNA extracted from indicated positions. Uncropped gels can be found in Supplementary Fig. 1. kDa, kilodaltons; NRS, non-immune serum; UV, ultraviolet. **b**, Size distribution and 5′ end nucleotide (nt) composition per size of CLIP tag. Error bars represent s.d.; $n=3$ (biological replicates, the same applies to **c**, **e** and **g**). **c**, Genomic distribution of CLIP tags for three high *yw* embryo (0–2 h) Aub CLIPs. **d**, Position of 5′ ends of retrotransposon lgCLIPs relative to 5′ ends of complementary piRNAs (0, *x* axis). **e**, Nucleotide composition at +9 of retrotransposon-derived lgCLIPs with 10-nucleotide overlap to complementary piRNAs. **f**, *yw* ovary Aub lgCLIP 5′ end positions relative to the 5′ ends of Ago3-loaded piRNAs (0, *x* axis). **g**, Schematic of processing fragments captured by Aub CLIP.

and Extended Data Fig. 5). Four separate analyses provide strong evidence that the extent of the observed Aub binding of mRNAs cannot be explained by piRNA targeting of transposon sequences embedded in mRNAs (Supplementary Results and Extended Data Fig. 6).

To investigate the potential of piRNAs to direct Aub to complementary mRNA sequences further, we analysed chimaeric lgCLIPs[23,24] that each contain an intact piRNA, ligated with a sequence fragment (≥20 nucleotides) that is uniquely aligned on mRNAs (Fig. 2a and Supplementary Table 5). To uncover complementarity patterns, we implemented unweighted local alignment between the piRNA (in reverse complement orientation) and the mRNA fragment, scoring matches (+1), mismatches (−1) and indels (−2), and reporting the best alignment for every chimaeric read. The search was performed within ±100 bases around the midpoint of the mRNA fragment; this allows the identification of the entire complementary sequence that might be missing from the chimaeric fragment, and also provides a reliable estimate of the signal-to-noise ratio. We observed prominent peaks of hundreds of thousands of complementarity events forming around the midpoint and within ±25 nucleotides, in *yw* and *tud* embryo CLIP libraries (Fig. 2b, c). Most events score between 7 and 12; therefore, the

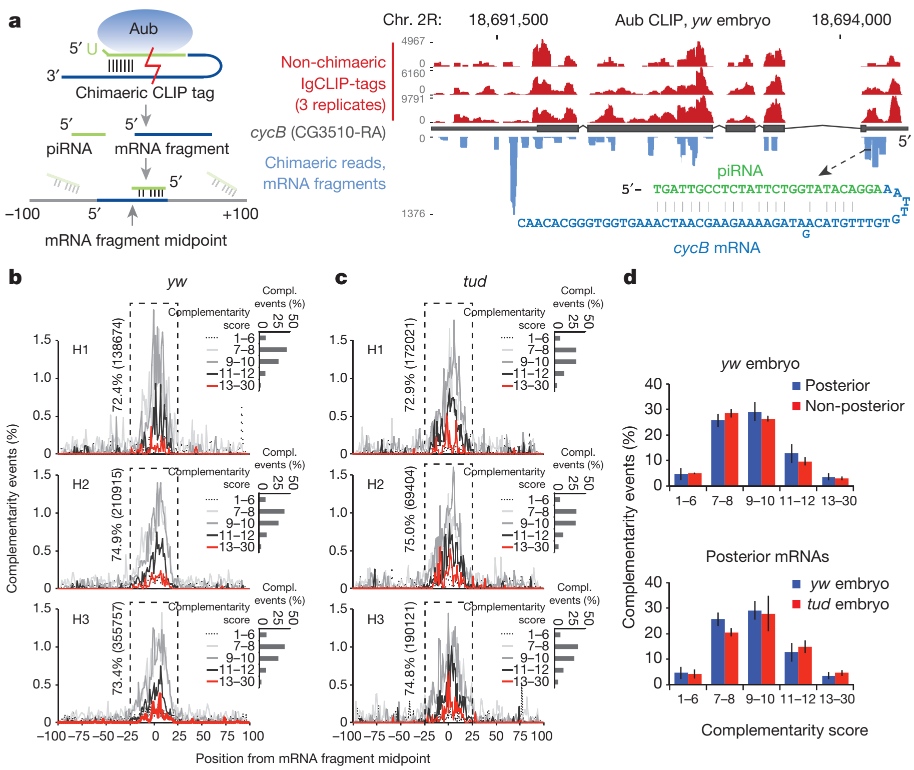


Figure 2 | Complementarity analysis between the piRNA and mRNA parts of chimaeric CLIP tags. **a**, Strategy for chimaeric CLIP tag analysis, and genome browser illustrating Aub lgCLIPs on *cycB*; sequence and base pairing of a chimaeric CLIP tag is shown. Chr, chromosome. **b**, **c**, piRNA–mRNA complementarity events (percentage) within ±100 bases from the midpoint of the mRNA part of the chimaeric read, plotted per alignment score for *yw* (**b**) and *tud* (**c**) embryo Aub CLIPs (biological triplicates). Percentage and number of total events occurring within ±25 bases (dashed rectangles) are shown. Inset, per sample: bar chart of number of complementarity events per score group. **d**, Bar charts of average piRNA–mRNA complementarity events occurring within the ±25-base window of the midpoint of the search area for indicated scores and mRNA localization categories and Aub CLIP libraries. All error bars denote s.d.; $n=3$.

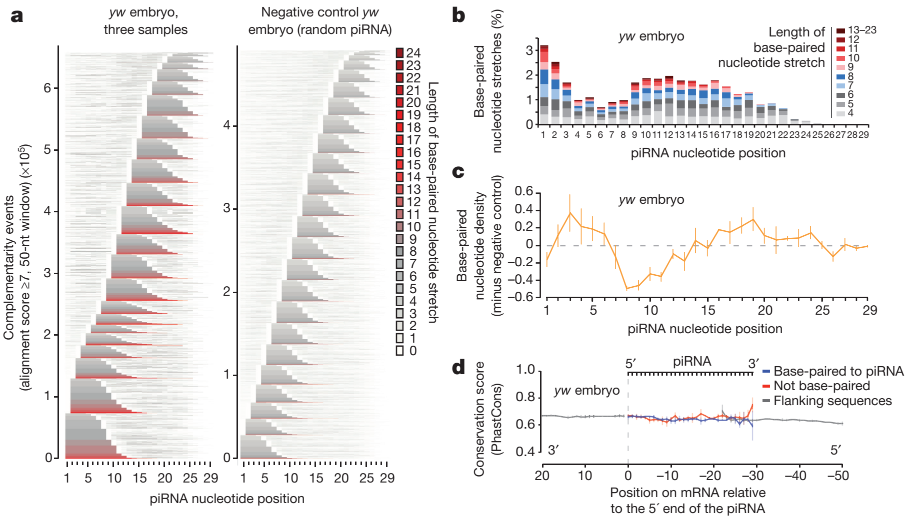


Figure 3 | Characteristics of piRNA base-pairing identified by chimaeric CLIP tag analysis. a, Heat maps showing base-paired nucleotides within the piRNA sequence, for all complementarity events (score ≥7) within ±25-base window, for *yw* embryo and negative control. Stacked piRNAs are sorted (bottom to top) by: starting position and length of the longest stretch and total number of base-paired nucleotides. Every nucleotide position is coloured according to the length of the stretch of consecutively base-paired nucleotides that runs through that position. **b**, Percentage of stretches of consecutive base-paired residues per starting position within the piRNA sequence. **c**, Average base-paired nucleotide density per position minus negative control (random piRNA). **d**, Average mRNA conservation score on and around piRNA–mRNA contact sites. All error bars denote s.d.; $n = 3$.

complementarity is not extensive. The distribution of the complementarity events in the negative control (random piRNA) is completely flat across the search area and has lower scores (Extended Data Fig. 7a), suggesting that the chimaeric reads capture genuine sequence-dependent Aub-piRNA–mRNA contacts.

piRNAs in chimaeric reads are typical Aub piRNAs (Extended Data Fig. 7b–e). piRNA–mRNA complementarities with alignment score ≥7 congregate within a 50-nucleotide window (Fig. 2b–d), so we focused on events that have such scores and locations. piRNA complementarity towards posterior and non-posterior mRNAs is indistinguishable (Fig. 2d and Extended Data Fig. 7f), suggesting that the basis of mRNA binding preference by Aub is not sequence specificity. Chimaeric reads show substantial overlap (Fig. 2a) and the same enrichment in posterior-localized mRNAs with non-chimaeric lgCLIPs (Supplementary Tables 5 and 6), suggesting that they both capture the same RNA binding events.

Base-paired nucleotides for every piRNA from three replicate CLIP libraries are summarized in a comprehensive plot (Fig. 3a and Extended Data Fig. 7g), revealing a bimodal distribution of the complementary regions within the piRNA. Many are found at the 5′ end of the piRNA, starting at positions 1 and 2 (reminiscent of miRNA seed-type binding); additional base-paired stretches start at positions 9–17 (Fig. 3a, b). This pattern is absent from the negative control (Fig. 3a). Net density of base-paired nucleotides reveals a clear preference for piRNAs to use nucleotides at positions 2–6 with additional base pairs in positions 16–24 (Fig. 3c and Extended Data Fig. 7h, i). This profile is markedly similar in *yw* and *tud* libraries, and differs slightly from the miRNA hybridization profile[24] in the less frequent base-pairing in the 2–6 region, suggesting that piRNAs do not use a conserved seed sequence. The periodicity of the graph in Fig. 3c (Extended Data Fig. 7i) evokes the helical conformation and base-pairing availability of the small RNA in the context of an Ago–miRNA–target RNA tripartite complex[25], suggesting that despite the absence of a conserved seed, the mechanics of piRNA complementary binding are analogous to those of microRNAs. Analysis of the evolutionary conservation of paired, unpaired and flanking nucleotides on the mRNA sequence reveals that the piRNA–mRNA contact sites are not preferentially conserved (Fig. 3d).

We used the local alignment approach by which we analysed the chimaeric CLIP tags, to identify potential piRNA target sites in the *D. melanogaster* transcriptome. In 206,400,271 total sites, the vast majority (99.6%) are of scores 7–11 (Fig. 4a). Importantly, the densities of putative piRNA target sites on mRNA regions are essentially identical for mRNAs with or without posterior localization, and very similar to that of the chimaeric mRNA fragments (higher densities in the UTRs compared to the coding sequences; Fig. 4b, c and Extended Data Fig. 8).

mRNAs in the 12 posterior localization categories are significantly longer than non-posterior localized mRNAs[26] (Fig. 4d), and so contain a higher number of piRNA target sites (Fig. 4e); nevertheless, transcript length normalization eliminates this difference (Fig. 4f, g). This holds true when the scores of the predicted sites are accounted for (Fig. 4g), and also when the scores are weighted for the preference of piRNA nucleotides 2–6 and 16–24 to base-pair (not shown). Posterior mRNAs are also more abundant than non-posterior; when factored in, this increases the difference of the target site abundance per transcript for the two localization categories (Fig. 4h). Posterior and non-posterior mRNAs are equally targeted (per kilobase) by each piRNA even when piRNA copy number is accounted for (Extended Data Fig. 9a). Notably, the size differential (and not the absolute length) of posterior and non-posterior mRNAs is conserved among drosophilids: the intra-species size differential always favours posterior mRNAs, although non-posterior mRNAs from one species might be longer than the posterior mRNAs of another (Fig. 4i). Therefore, although piRNAs randomly base pair with non-conserved mRNA sequences, this mechanism is biased towards a specific class of mRNAs for germ plasm anchoring. Additionally, from the two categories of posterior localized mRNAs, 'localized' and 'protected'[10], localized mRNAs have longer 3′ UTRs than protected mRNAs, further supporting the notion that mRNA length positively affects germ plasm enrichment (Extended Data Fig. 9b, c).

The concept of mRNA entrapment at the germ plasm during ooplasmic streaming is well established[8,9,27], but the mechanism at the molecular level has so far been elusive. We propose that germ plasm localized Tud–Aub–piRNA complexes play the role of a non-discriminatory adhesive trap that can form numerous, non-conserved piRNA–mRNA contacts to capture mRNAs and form germ plasm mRNPs (Fig. 4j and Supplementary Discussion). This mechanism probably shows preference for posterior mRNAs because they are

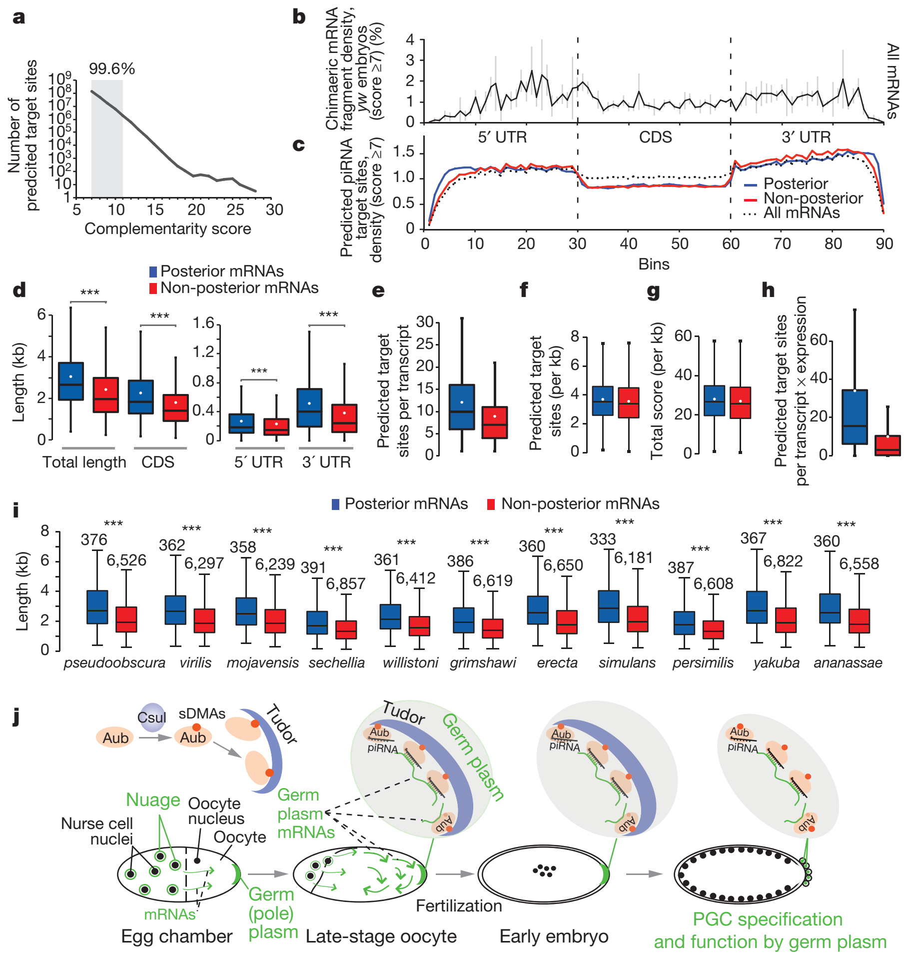


Figure 4 | Transcriptome-wide prediction of piRNA target sites and length differential of posterior-localized mRNAs. a, Number of predicted piRNA complementary sites on mRNAs, per score. **b**, **c**, Average binned density of: chimaeric mRNA fragments (Aub CLIP, *yw* embryo 0–2 h) along the meta-mRNA. CDS, coding sequence. Error bars denote s.d.; $n = 3$ (**b**); predicted piRNA complementary sites within all (14,058), posterior (380), and non-posterior (6,747) localized mRNAs (**c**). **d**–**i**, Box-and-whisker plots of: lengths of mRNAs expressed in *yw* embryos (0–2 h) (**d**); number of predicted piRNA complementary sites per mRNA (**e**); length-normalized number of predicted piRNA complementary sites (**f**); length-normalized total score of predicted piRNA complementary sites (**g**); number of predicted piRNA complementary sites per mRNA multiplied by the abundance of each mRNA RPKM (reads per kilobase per million mapped reads) (**h**); and lengths of orthologous mRNAs in other *Drosophila* species (**i**). Black lines denote median; white dots denote mean. $^{***}P < 0.005$, one-sided *t*-test (**d**); $^{***}P < 1 \times 10^{-16}$, one-sided Wilcoxon exact rank test (**i**). **j**, Aub couples piRNA inheritance with germ-cell specification in *Drosophila*. Aub, carrying symmetrically dimethylated arginine residues (sDMAs) dimethylated by Csul methyltransferase, interacts with Tudor, and both are localized in the germ plasm during mid-stage oogenesis. Ooplasmic streaming at later stages promotes diffusion of mRNPs, facilitating random contacts of mRNAs with the germ plasm. Aub piRNAs form an adhesive trap that captures mRNAs forming numerous low complementarity contacts. mRNAs with posterior functions are longer and more abundant than the rest, form more piRNA-mediated contacts with the germ plasm, and thus their entrapment is enhanced. Tudor–Aub-piRNA–mRNA complexes along with other RNA binding proteins form germ granules that contain both piRNAs and mRNAs that induce primordial germ cell (PGC) specification. Aub and its RNA cargo is incorporated in PGCs, providing the maternal mRNAs that are necessary for PGC function and the maternal piRNAs that will propagate an RNA immune response against transposons.

significantly longer and more abundant[26]. We believe that the above mechanism acts in addition to specific protein–protein, protein–RNA and RNA–RNA interactions that are necessary for mRNA transfer and anchoring to the posterior, and for translational control[10,12,28–30]. The multivalence of Aub–Tudor interactions probably contributes to the formation of multimeric germ granule complexes. We propose that germ-cell specification and function by maternal mRNAs, and piRNA inheritance converge in Aub. Coupling germ-cell specification with piRNA inheritance could be a strategy that increases reproductive fitness by ensuring the propagation of robust transposon silencing mechanisms to germ cells across generations and across the population.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 September 2015; accepted 13 January 2016.
Published online 7 March 2016.

1. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Rev. Mol. Cell Biol.* **12,** 246–258 (2011).
2. Ephrussi, A. & Lehmann, R. Induction of germ cell formation by oskar. *Nature* **358,** 387–392 (1992).
3. Mahowald, A. P. Assembly of the *Drosophila* germ plasm. *Int. Rev. Cytol.* **203,** 187–213 (2001).
4. Brennecke, J. *et al.* An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322,** 1387–1392 (2008).
5. Grentzinger, T. *et al.* piRNA-mediated transgenerational inheritance of an acquired trait. *Genome Res.* **22,** 1877–1888 (2012).
6. Khurana, J. S. *et al.* Adaptation to P element transposon invasion in *Drosophila melanogaster*. *Cell* **147,** 1551–1563 (2011).
7. Bucheton, A. Non-Mendelian female sterility in *Drosophila melanogaster*: influence of aging and thermic treatments. III. Cumulative effects induced by these factors. *Genetics* **93,** 131–142 (1979).

8. Kugler, J. M. & Lasko, P. Localization, anchoring and translational control of oskar, gurken, bicoid and nanos mRNA during *Drosophila* oogenesis. *Fly (Austin)* **3,** 15–28 (2009).
9. Forrest, K. M. & Gavis, E. R. Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for *nanos* mRNA localization in *Drosophila*. *Curr. Biol.* **13,** 1159–1168 (2003).
10. Rangan, P. *et al.* Temporal and spatial control of germ-plasm RNAs. *Curr. Biol.* **19,** 72–77 (2009).
11. Thomson, T., Liu, N., Arkov, A., Lehmann, R. & Lasko, P. Isolation of new polar granule components in *Drosophila* reveals P body and ER associated proteins. *Mech. Dev.* **125,** 865–873 (2008).
12. Trcek, T. *et al. Drosophila* germ granules are structured and contain homotypic mRNA clusters. *Nature Commun*. **6,** 7962 (2015).
13. Kirino, Y. *et al.* Arginine methylation of Aubergine mediates Tudor binding and germ plasm localization. *RNA* **16,** 70–78 (2010).
14. Liu, H. *et al.* Structural basis for methylarginine-dependent recognition of Aubergine by Tudor. *Genes Dev.* **24,** 1876–1881 (2010).
15. Arkov, A. L., Wang, J.-Y. S., Ramos, A. & Lehmann, R. The role of Tudor domains in germline development and polar granule architecture. *Development* **133,** 4053–4062 (2006).
16. Boswell, R. E. & Mahowald, A. P. *tudor*, a gene required for assembly of the germ plasm in *Drosophila melanogaster*. *Cell* **43,** 97–104 (1985).
17. Vourekas, A. *et al.* Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. *Nature Struct. Mol. Biol*. **19,** 773–781 (2012).
18. Mohn, F., Handler, D. & Brennecke, J. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science* **348,** 812–817 (2015).
19. Lécuyer, E. *et al.* Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **131,** 174–187 (2007).
20. Thomson, T. & Lasko, P. *Drosophila tudor* is essential for polar granule assembly and pole cell specification, but not for posterior patterning. *Genesis* **40,** 164–170 (2004).
21. Barckmann, B. *et al.* Aubergine iCLIP reveals piRNA-dependent decay of mRNAs involved in germ cell development in the early embryo. *Cell Rep.* **12,** 1205–1216 (2015).
22. Rouget, C. *et al.* Maternal mRNA deadenylation and decay by the piRNA pathway in the early *Drosophila* embryo. *Nature* **467,** 1128–1132 (2010).
23. Moore, M. J. *et al.* miRNA-target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nature Commun*. **6,** 8864 (2015).
24. Grosswendt, S. *et al.* Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Mol. Cell* **54,** 1042–1054 (2014).
25. Schirle, N. T., Sheu-Gruttadauria, J. & MacRae, I. J. Structural basis for microRNA targeting. *Science* **346,** 608–613 (2014).
26. Jambor, H. *et al.* Systematic imaging reveals features and changing localization of mRNAs in *Drosophila* development. *eLife* **4,** e05003 (2015).
27. Sinsimer, K. S., Lee, J. J., Thiberge, S. Y. & Gavis, E. R. Germ plasm anchoring is a dynamic state that requires persistent trafficking. *Cell Rep*. **5,** 1169–1177 (2013).
28. Little, S. C., Sinsimer, K. S., Lee, J. J., Wieschaus, E. F. & Gavis, E. R. Independent and coordinate trafficking of *Drosophila* germ plasm mRNAs. *Nature Cell Biol.* **17,** 558–568 (2015).
29. Ghosh, S., Marchand, V., Gáspár, I. & Ephrussi, A. Control of RNP motility and localization by a splicing-dependent structure in oskar mRNA. *Nature Struct. Mol. Biol*. **19,** 441–449 (2012).
30. Gavis, E. R., Lunsford, L., Bergsten, S. E. & Lehmann, R. A conserved 90 nucleotide element mediates translational repression of nanos RNA. *Development* **122,** 2791–2800 (1996).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank former and current laboratory members for discussions; M. Siomi for the Tudor antibody; A. Arkov for *tud* flies; G. Dreyfuss for the PABP antibody; and J. Schug for Illumina sequencing. Work was supported by a Brody family fellowship to M.M., and a National Institutes of Health (NIH) grant GM072777 to Z.M.

Author Contributions A.V. and Z.M. conceived, and Z.M. supervised, the study. A.V. and N.V. performed the experiments. P.A. performed bioinformatic analyses with contribution from M.M. and A.V. A.V., P.A., N.V., M.M. and Z.M. interpreted the data. A.V. wrote the manuscript, with contribution from all authors.

Author Information Sequences have been deposited in the Sequence Read Archive (SRA) under the accession number SRP067739. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.M. (mourelaz@uphs.upenn.edu).

METHODS

***Drosophila* strains, tissue collection.** The following strains and heteroallelic combinations were used: $y^{1}w^{1118}$ as the wild-type stock (*yw*), $aub^{HN2/QC42}$ (*aub*) and $tud^{1/Df(2R)PurP133}$ (*tud*), for *aub* and *tud* mutants (loss-of-function), respectively[15,31–33]. All flies were grown at 25 °C with 70% relative humidity on a 12-h light–dark cycle. The 2–4-day female flies were crossed to *yw* males for 2 days in standard cornmeal food supplied with yeast paste before ovary dissection. Embryos collected at well-defined time-windows were dechorionated in 50% commercial bleach for 2 min, washed extensively in water and collected in PBS or HBSS or fixation solution, depending on downstream applications.

Antibodies. Antibody against Aubergine (Aub-83) was produced by immunizing rabbits with Aub peptide (HKSEGDPRGSVRGRC, in which terminal cysteine was used to couple to KLH; Genscript) and selected with peptide-affinity purification of sera. Other antibodies that were used in this study: mouse monoclonal anti-PABP (6E2 clone)[34], E7 mouse monoclonal anti-β-tubulin (Developmental Studies Hybridoma Bank) and anti-Tudor mouse monoclonal (gift from M. Siomi).

Immunofluorescence. Fixation and immunohistochemistry of dissected ovaries and embryos was performed according to standard protocols. Primary antibodies against Aub and Tud were used at 1 ng μl^{-1} final concentration. Secondary antibodies conjugated to Alexa 488 and 594 (Life technologies) were used at 1:1,000 dilution. Ovary and embryo samples were imaged on Leica TCS SPE confocal microscope.

Aub CLIP-seq (HITS-CLIP, high-throughput sequencing after crosslinking and immunoprecipitation). CLIP was performed as previously described for Mili, Miwi and MOV10L1 (refs 17, 35, 36). The protocol is described in detail previously[36] and uses stringent buffer conditions to ensure high specificity. The experiment was performed in three biological replicates for each condition (*yw* ovaries, *yw* embryos 0–2 h, *tud* embryos 0–2 h). Approximately 40 mg of *Drosophila* embryos (0–2 h) or ~80 ovaries from 4–6-day females were collected in ice-cold HBSS and ultraviolet-irradiated (3×) at 254 nm (400 mJ cm^{-2}). The tissues were pelleted, washed with PBS and the final tissue pellet was flash-frozen in liquid nitrogen and kept at −80 °C. Ultraviolet-light-treated tissues were lysed in 350 μl PMPG (PBS (no Mg^{2+} and no Ca^{2+}), 2% Empigen) with protease inhibitors and rRNasin (2 U μl^{-1}) and no exogenous RNases; lysates were treated with DNase I (Promega) for 5 min at 37 °C, and then were centrifuged at 100,000*g* for 30 min at 4 °C.

For each immunoprecipitation, approximately 10 μl of our anti-Aub antibody was bound on 150 μl (slurry) of protein A Dynabeads in Ab binding buffer (0.1 M Na-phosphate, pH 8, and 0.1% NP-40) at room temperature for 2 h; antibody-bound beads were washed three times with PMPG. Antibody beads were incubated with lysates (supernatant of 100,000*g*) for 3 h at 4 °C. Low- and high-salt washes of immunoprecipitation beads were performed with 1× and 5× PMPG (5× PBS, 2% Empigen). RNA linkers (RL3 and RL5), as well as 3′ adaptor labelling and ligation to CIP (calf intestinal phosphatase)-treated RNA CLIP tags were performed as previously described[36].

Immunoprecipitation beads were eluted at 70 °C for 12 min using 30 μl of 2× Novex reducing loading buffer. Samples were analysed by NuPAGE (4–12% gradient precast gels, run with MOPS buffer). Cross-linked RNA–protein complexes were transferred onto nitrocellulose (Invitrogen), and the membrane was exposed to film for 1–2 h. Membrane fragments containing the main radioactive signal and fragments up to ~15 kDa higher were excised (Fig. 1a). RNA extraction, 5′ linker ligation, Reverse transcriptase PCR (RT–PCR) and a second PCR step were performed with the DNA primers (DP3 and DP5, DSFP3 and DSFP5) as described previously[36]. Complementary DNA from two PCR steps was resolved on and extracted from 3% Metaphor 1× TAE gels. Size profiles of cDNA libraries prepared from the main radioactive signal and higher molecular mass signal were similar (Fig. 1a). DNA was extracted with QIAquick Gel Extraction kit and submitted for deep sequencing. The cDNA libraries were sequenced with Hi-Seq Illumina at 100 cycles.

Solid-support directional RNA-seq. Solid-support directional RNA-seq was performed as previously described[17], using total RNA (depleted of ribosomal RNA with Ribo-Zero (EpiCentre)) isolated from 0–2-h embryos of appropriate genotypes.

Nycodenz density gradient ultracentrifugation and subsequent analyses. Nycodenz density gradient separation of RNPs was performed as previously described[17] with modifications. A 20–60% (top to bottom) Nycodenz gradient (4.8 ml) in 1× KMH150 (150 mM KCl, 2 mM $MgCl_2$, 20 mM HEPES, pH 7.4, 0.5% NP-40, 0.1 U μl^{-1} rRNasin, and protease inhibitors) was prepared as a step gradient by overlaying five equal parts of Nycodenz solutions and was left to diffuse overnight at 4 °C. 0.2 microlitres of post-nuclear *yw* embryo lysate in 1× KMH150 was laid over the gradient and centrifuged at 150,000*g* for 20 h. We used embryos of stages 4–6, to avoid earlier stages where mRNAs at the soma form distinct mRNPs than the ones formed in the pole plasm PGCs. The gradient was collected in 12 equal fractions. Samples from each fraction were used for protein determination by Bradford and RNA extraction with Trizol LS. Right before RNA extraction, 500 ng of *in vitro* transcript of *Renilla* luciferase mRNA was spiked in each fraction for normalization purposes in subsequent steps.

qRT–PCR. An equal volume of RNA extracted from each fraction was reverse transcribed by Supersript III (Invitrogen 18080-051) in the presence of random hexamers. Equal volume of the cDNA was mixed with primers (*gcl*, *osk*, *Hsp83*, *dhd*, *cycB*: Qiagen QuantiTect Assay; *Renilla* luciferase (rLuc), forward: 5′-CGCTGAAAGTGTAGTAGATGTG-3′ and reverse: 5′-TCCACGAAGAAGTTATTCTCCA-3′) and Power SYBR Green reaction mix (Applied Biosystems 4367659). The reactions were run on a StepOnePlus System (Applied Biosystems) using the default program.

Immunoprecipitation and detection of piRNAs, and preparation of cDNA libraries. Aub immunoprecipitation, 5′ end labelling of piRNAs and cDNA library preparation were carried out as previously described[37,38].

Code availability. We used CLIPSeqTools[39], a bioinformatics suite that we created for analysis of CLIP-seq data sets (accessible at: http://mourelatos.med.upenn.edu/clipseqtools/) and a Perl programming framework that we developed[40]. The latter framework is named GenOO and has been specifically developed for analysis of high-throughput sequencing data. The source code for GenOO has been deposited in GitHub and can be accessed at https://github.com/genoo/.

Statistics. In statistical analyses, we ensured that the assumptions of each statistical test are met and that the statistical test used is appropriate for the analysis. In all analyses the statistical tests and methods used are clearly stated in relevant sections. No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Data. *Drosophila* (assembly dm3) transcript, exon and repeat genomic locations were downloaded from the UCSC genome browser (downloaded 22 March 2011 from http://genome.ucsc.edu). Repeat consensus sequences were downloaded from Flybase (http://flybase.org/ - transposon_sequence_set v9.42). Localization categories for *Drosophila* genes were taken from ref. 19. The localization annotation matrix was downloaded from (http://fly-fish.ccbr.utoronto.ca/annotation_matrix.csv). Transposon categories were as in ref. 31.

Preprocessing. The 3′ end ligated adaptor (GTGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCTGCTTG) was removed from the sequences using the cutadapt software and a 0.25 acceptable error rate for the alignment of the adaptor on the read. To eliminate reads in which the adaptor was ligated more than one time, adaptor removal was performed three times.

Alignment. Reads for all samples were aligned against the dm3 *Drosophila melanogaster* genome assembly using the aligner bwa v0.6.2-r126, with the default settings[41]. Reads were also aligned against the Repeat consensus sequences using the same aligner.

Genomic distribution. All mapped reads were divided in the following genomic categories: sense repeat, antisense repeat, non-coding RNA, (protein) coding RNA. The remaining reads were considered to be intergenic reads.

Correlation of replicates. Gene expression was defined as the number of reads that map on each gene and the values were normalized by the upper quartile normalization method[42]. The $\log_2$ gene expression levels of replicates are compared using the Pearson Correlation function in R.

Coincidence with immunoprecipitation. Reads mapping in the same position (same 5′ end mapping) were considered as coinciding. When comparing CLIP with immunoprecipitation libraries, the percentage of piRNA-size CLIP reads that had a coinciding start with any standard immunoprecipitation read were counted as positive.

Significant localization. For each localization category, the quartile-normalized lgCLIP binding level ('mRNA expression level' in each CLIP library) is compared via a two-sided *t*-test between genes that belong to the category versus genes that do not belong to it. To compare two samples, we measure the difference in binding (per gene) between the two conditions ($\log_2$(gene.expr.cond1/gene.expr.cond2)) and then perform a *t*-test of differences in genes belonging to the category versus genes not belonging in the category.

Early embryo posterior localization categories. The following twelve mRNA localization categories[19] were found significantly depleted in *tud* embryo Aub CLIP libraries compared to *yw* embryo libraries, and were used in analyses were 'posterior localized mRNAs' are mentioned: '1:41:RNA islands', '1:42:Pole buds', '1:40:Pole plasm', '3:265:Perinuclear around pole cell nuclei', '4:370:Germ cell localization', '4:403:Germ cell enrichment', '3:348:Pole cell enrichment', '2:141:Pole cell localization', '2:153:Perinuclear around pole cell nuclei', '2:142:Pole cell enrichment', '3:347:Pole cell localization', '1:59:Perinuclear around pole cell nuclei'

(http://fly-fish.ccbr.utoronto.ca/). The remaining mRNAs are mentioned as non-posterior localized mRNAs. The following three posterior localization categories were also depleted in *tud* embryo Aub CLIP libraries compared to *yw*: '1:39:Posterior localization', '2:124:Posterior localization', '3:352:Posterior localization'. Almost all of the mRNAs contained in the above twelve categories are also contained in these three, but these three categories also contain some mRNAs that do not actually localize in the pole plasm or the germ cells (that is, with apical localization); therefore, mRNAs belonging in any of these three localization categories but not in any of the above mentioned twelve posterior categories were not considered for the generation of the Supplementary Table 4. Many mRNAs do not have a designated localization pattern, and they are mentioned as 'undetermined localization'. It is worth mentioning that this category contains several mRNAs with clear posterior–pole-plasm localization. Through manual searches of the Berkeley *Drosophila* Genome Project chromogenic ISH database (http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl) we noticed that many Aub-bound mRNAs, the localization of which is not annotated in the Fly-FISH database, are indeed localized in the germ plasm/cells (such as *CG4735/shu*, *CG7070/PyK*, *CG4903/MESR4*, *CG5452/dnk* and *CG9429/Calr*), therefore our analysis is most likely underestimating the true number of Aub-bound mRNAs that are important for germline specification and function. Because of this, mRNAs with 'undetermined localization' were never mixed with 'non-posterior localized' mRNAs in our analyses.

Highly bound genes. To identify highly bound genes, we used the rank product method[43]. Specifically, genes are sorted by expression per sample, and for each gene the product of their ranks is calculated. The probability of this rank product produced by chance is calculated by permutations of all non-zero value genes.

Transcript expression calculation. We calculated the expression for protein-coding transcripts by counting the number of RNA-seq reads that map within the exons of each transcript. The counts were normalized using RPKM and upper quartile normalization, effectively dividing each count by the upper quartile of all counts[42]. The transcript with the highest RPKM score was used ('best transcript') unless otherwise noted.

Transcript Aub-binding calculation. We calculated the expression for protein-coding transcripts by counting the number of CLIP reads that map within the exons of each transcript in the sense orientation. The counts were normalized using reads per million and upper quartile normalization, effectively dividing each count by the upper quartile of all counts[42].

RNA-seq correlation versus CLIP. Upper quartile normalized RPKM for RNA-seq was compared to similarly normalized CLIP binding levels defined as average number of reads per transcript in CLIP replicates. Correlation was calculated using the Pearson Correlation function in R.

Identification of hybrid reads. (1) Identified lgCLIP size reads (read length >35) that did not align to the genome. (2) Made a set of substrings from both ends of reads from (1) of piRNA size ($L=[23,29]$). (3) Identified the substring from (2) to full-length piRNAs ($L=[23,29]$) from corresponding low samples (Extended Data Fig. 1b) (4) The longest aligning piRNAs are retained and coupled with the remainder of the read as piRNA–lgCLIP couples. (5) The piRNA aligning fragment is cut from the read. Very small remainder reads ($L=[<20]$) are discarded. (6) The remainders are aligned to the genome (using bwa default settings). (7) Remainders aligned in one single position that is on a known mRNA are retained.

Alignment of piRNAs to regions. (1) Regions of 200-nucleotide length were cut around the midpoint of the genomic alignment region from step 7 of previous routine. Specifically, if ($d=200$ the length of the final region we want and L is the length of the read), a genomic region flanking the read on each side of length $d/2$ was excised from the chromosome sequence. If the alignment was located in the minus strand the sequence was reversed and complemented at this point. This total region has length $d+L$. We discard an equal number of nucleotides from each side to reach a final length of L (specifically we substring starting from int($L/2$) and for d nucleotides. Note, int will always round down). At this point we have a region of length 200 nucleotides centred around the alignment region of the fragment. (2) We use a slightly modified Smith–Waterman[44] alignment method (weights: match $=+1$, mismatch $=-1$, gap $=-2$) to align piRNAs on the 200-nucleotide long regions from (1). Differences of our alignment versus Smith–Waterman: (a) No penalties are given to non-matching nucleotides on the edges of the alignment. (b) If there are multiple optimal alignment scores, one is picked randomly. (c) Alignments in which part of one sequence is outside the boundaries of the other sequence are not considered. (3) The midpoint of the alignment (if k nucleotides matched that is the int($k/2$) nucleotide) is used for graphs of alignment positioning on regions.

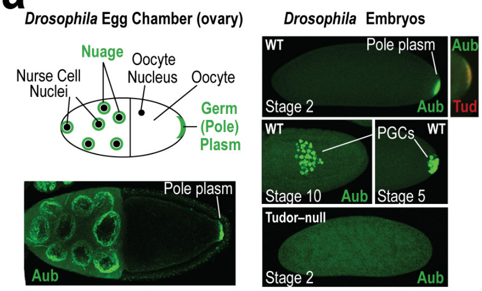
mRNA target prediction for the top 2,000 expressed piRNAs. We grouped piRNA sequences into families based on the first 23 nucleotides of each piRNA. Using the alignment algorithm described above we aligned one piRNA (the most abundant) for each of the top 2,000 families to the longest annotated transcript for each protein-coding gene. These 2,000 piRNA families represent ~37% of piRNA reads from low *yw* CLIP libraries. To factor in transcript abundance, we multiplied the RNA-seq (*yw* 0–2-h embryo) RPKM value for each mRNA with the number of predicted piRNA target sites found within the mRNA. This provides a 'targeting potential' of every mRNA species, corrected for its abundance.

We then evaluated the targeting potential of each piRNA–mRNA pair using three different scoring schemes. For the first, we sum the alignment score of all putative piRNA binding sites on the mRNA. For the second, we calculated a weighted alignment score for each putative piRNA binding site and then we sum all scores similar to the previous scheme. The weighted score for each binding site is calculated based on the following formula $\sum_i x_i * A_i$, in which x_i is 1 or 0 based on whether the nucleotide at position $\boldsymbol{i}$ of the piRNA is bound or not, and A_i is the weight for nucleotide $\boldsymbol{i}$. For the third, we multiplied the total number of predicted complementary sites per piRNA, with the piRNA copy number.

Study of the lengths of *D. melanogaster* orthologous mRNAs in other *Drosophila* species. Transcript sequences (fasta file) for each species were downloaded from Flybase (ftp://ftp.flybase.net/genomes/ on 1 September 2015, current version used for each genome). For each gene (identified as the 'parent' tag in the fasta file header), the longest transcript length was identified. For the analysis of the expressed mRNAs (Fig. 4d), we used our *yw* embryo RNA-seq data to identify the longest transcript with the highest length normalized abundance. Orthologue gene tables were downloaded from Flybase (gene_orthologs_fb_2015_03.tsv.gz) and were used to identify orthologue genes across species. For each species, all genes that mapped to localized and unlocalized *Drosophila melanogaster* genes were used in the comparison and were assigned to the corresponding group as their *D. melanogaster* orthologue. Boxplots were created using the lattice package in R (bwplot) and omitting outliers, P values were calculated using the Wilcoxon exact rank test (wilcox.test in R) one-sided with the hypothesis that localized genes are longer than non-localized.

31. Malone, C. D. *et al.* Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137,** 522–535 (2009).
32. Wilson, J. E., Connell, J. E. & Macdonald, P. M. *aubergine* enhances *oskar* translation in the *Drosophila* ovary. *Development* **122,** 1631–1639 (1996).
33. Schupbach, T. & Wieschaus, E. Female sterile mutations on the second chromosome of *Drosophila melanogaster*. II. Mutations blocking oogenesis or altering egg morphology. *Genetics* **129,** 1119–1136 (1991).
34. Matunis, M. J., Matunis, E. L. & Dreyfuss, G. Isolation of hnRNP complexes from *Drosophila melanogaster*. *J. Cell Biol.* **116,** 245–255 (1992).
35. Vourekas, A. *et al.* The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing. *Genes Dev.* **29,** 617–629 (2015).
36. Vourekas, A. & Mourelatos, Z. HITS-CLIP (CLIP-Seq) for mouse Piwi proteins. *Methods Mol. Biol.* **1093,** 73–95 (2014).
37. Kirino, Y., Vourekas, A., Khandros, E. & Mourelatos, Z. Immunoprecipitation of piRNPs and directional, next generation sequencing of piRNAs. *Methods Mol. Biol.* **725,** 281–293 (2011).
38. Kirino, Y. *et al.* Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability. *Nature Cell Biol.* **11,** 652–658 (2009).
39. Maragkakis, M., Alexiou, P., Nakaya, T. & Mourelatos, Z. CLIPSeqTools-a novel bioinformatics CLIP-seq analysis suite. *RNA* **22,** 1–9 (2016).
40. Maragkakis, M., Alexiou, P. & Mourelatos, Z. GenOO: a modern perl framework for high throughput sequencing analysis. Preprint at http://biorxiv.org/content/early/2015/11/03/019265 (2015).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
42. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* **11,** 94 (2010).
43. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **573,** 83–92 (2004).
44. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).

a

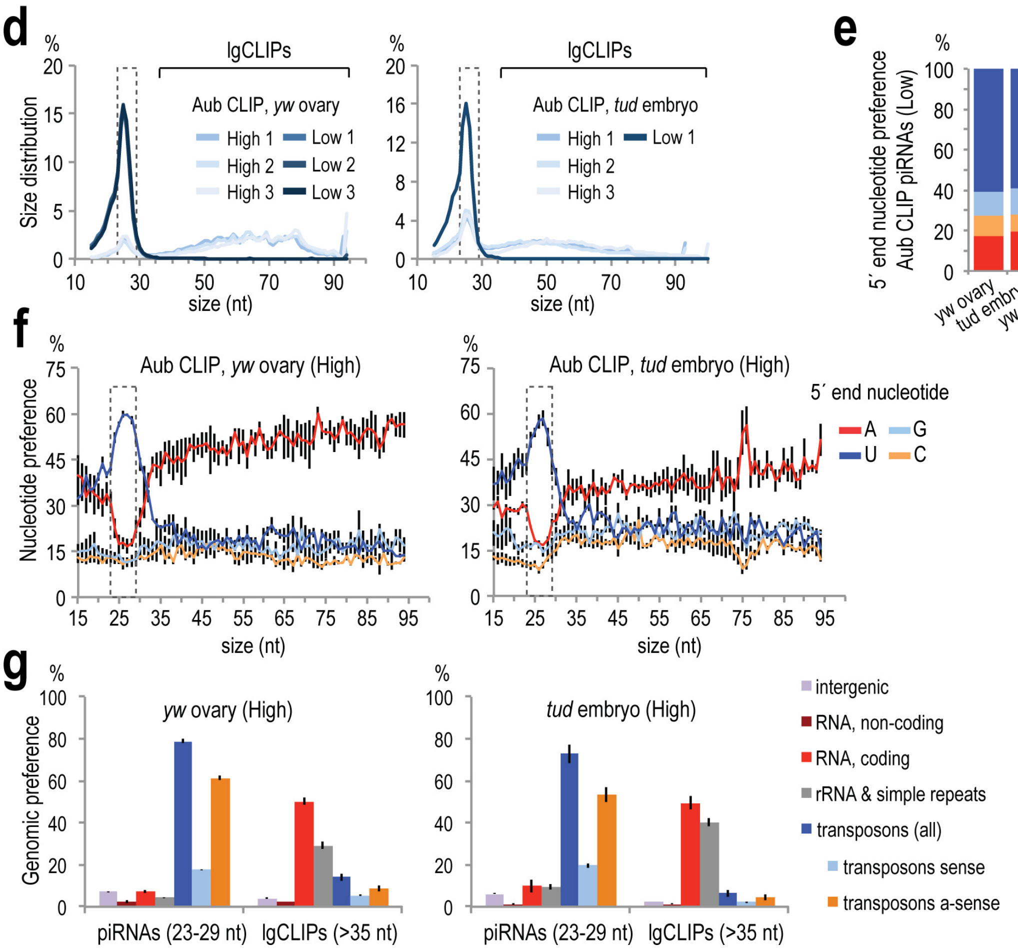


b

CLIP-Seq library	Number of sequenced reads	Number of mapped reads	piRNAs (23-29 nt)	lgCLIPs (>35 nt)
Aub_CLIP_yw_embryo_0_2_H_1	8,818,464	6,510,923	1,470,826	4,052,071
Aub_CLIP_yw_embryo_0_2_H_2	15,616,681	12,398,341	4,035,158	5,597,930
Aub_CLIP_yw_embryo_0_2_H_3	21,455,325	15,409,469	3,022,117	10,454,759
Aub_CLIP_yw_embryo_0_2_L_1	14,062,489	13,826,920	7,624,548	108,605
Aub_CLIP_yw_embryo_0_2_L_2	17,610,892	17,257,652	10,002,213	121,154
Aub_CLIP_yw_embryo_0_2_L_3	18,420,287	18,055,396	10,885,017	140,132
Aub_CLIP_yw_ovary_H_1	19,041,920	15,689,202	1,486,327	13,422,595
Aub_CLIP_yw_ovary_H_2	22,387,813	20,029,273	1,668,627	17,419,663
Aub_CLIP_yw_ovary_H_3	20,862,621	18,425,256	2,059,053	15,309,011
Aub_CLIP_yw_ovary_L_1	24,257,591	23,951,410	16,103,669	227,098
Aub_CLIP_yw_ovary_L_2	18,584,476	18,287,691	12,540,296	134,799
Aub_CLIP_yw_ovary_L_3	52,408,658	45,884,221	31,826,158	678,058
Aub_CLIP_tud_embryo_1_H	43,313,537	36,226,058	7,424,930	23,699,595
Aub_CLIP_tud_embryo_2_H	36,687,776	30,485,898	7,527,602	17,647,540
Aub_CLIP_tud_embryo_3_H	42,050,665	30,959,358	6,438,572	21,434,719
Aub_CLIP_tud_embryo_L	25,284,701	24,773,498	16,404,216	109,985

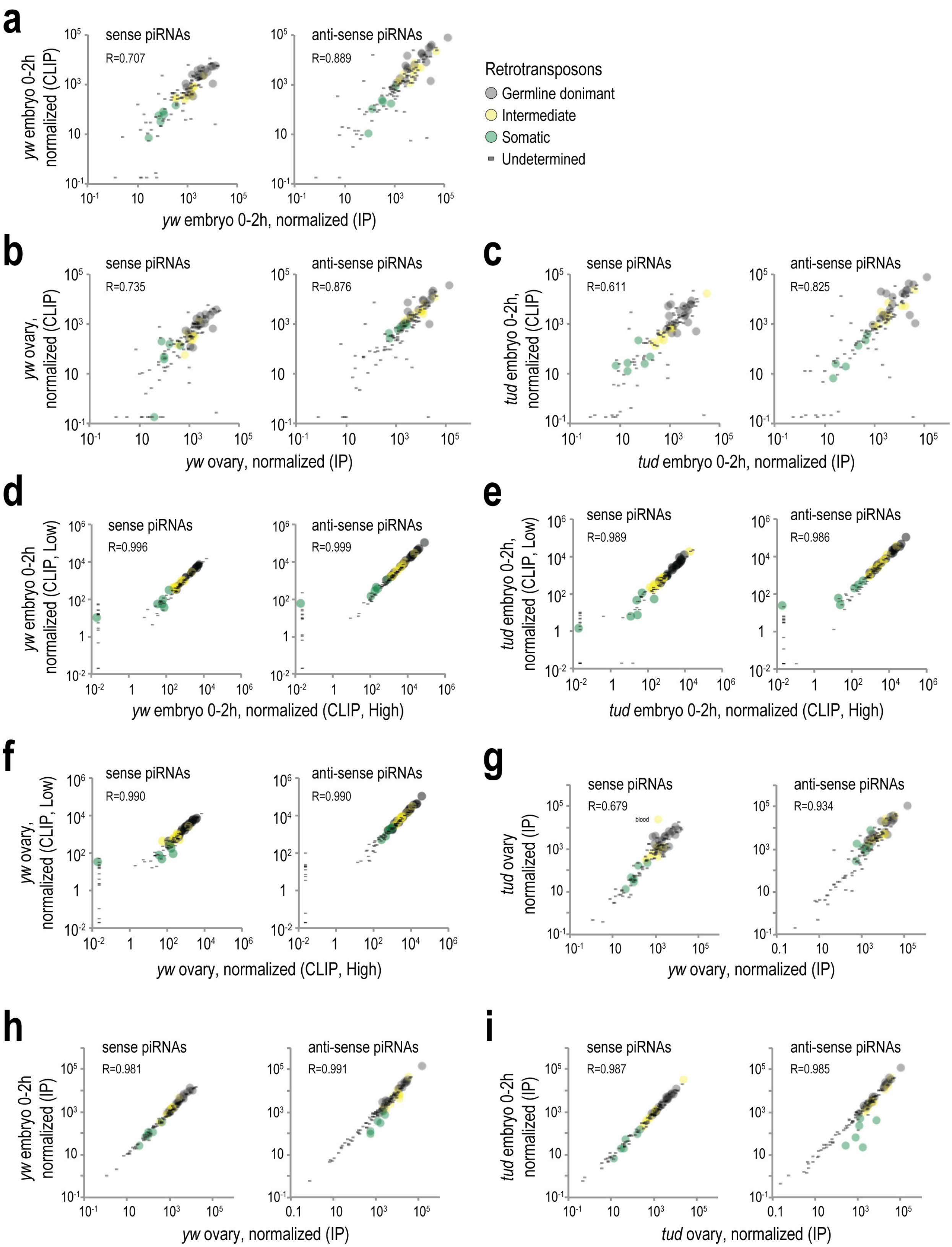
c

Small RNA-Seq library	Number of sequenced reads	Number of mapped reads
Aub_IP_tud_embryo_(0-2h)	35,614,928	35,368,281
Aub_IP_tud_embryo_(stage_4-6)	30,853,326	30,640,207
Aub_IP_tud_ovary	21,050,596	20,905,116
Aub_IP_yw_embryo_(0-2h)	43,405,719	42,958,773
Aub_IP_yw_embryo_(stage_4-6)	43,940,406	43,512,435
Aub_IP_yw_embryo_(stage_7-9)	35,047,292	34,694,141
Aub_IP_yw_ovary	71,829,805	70,295,208



Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | Endogenous Aub localization in genotypes used, sequenced and mapped reads of CLIP sequencing and RNA immunoprecipitation libraries used in this study, and general characteristics of *yw* ovary and *tud* embryo (0–2 h) CLIP sequencing libraries. a, Immunofluorescence of ovary and early embryo of indicated genotypes using antibodies against Aub (Aub-83; green) and Tudor (red), and schematic representation of the egg chamber. Aub is localized in the nuage and germ (pole) plasm of wild-type ovaries, in the germ plasm of early wild-type embryos (stage 2) and within PGCs as they form in the posterior pole (stage 5), and as they migrate during gastrulation (stage 10). Tudor colocalizes with Aub in the germ plasm of early embryos but is not detected after PGC formation. In Tudor mutant early embryos, Aub is not concentrated in the posterior but is diffusely present throughout the embryo; PGCs are never specified resulting in agametic adults (see also Extended Data Fig. 9). **b**, Sequenced and mapped reads of CLIP sequencing (CLIP-seq) libraries prepared in this study. **c**, Sequenced and mapped reads of RNA immunoprecipitation deep-sequencing libraries prepared in this study. **d**, Size distribution for the three low (one for *tud*) and three high *yw* ovary and *tud* embryo (0–2 h) Aub CLIP-seq libraries. The size range of piRNAs (23–29 nucleotides) is indicated by a dashed box. **e**, Average 5′ end nucleotide composition for piRNAs (23–29 nucleotides) from three low *yw* ovary, *tud* embryo (0–2 h) (one library) and *yw* embryo (0–2 h) Aub CLIP-seq libraries. **f**, Average 5′ end nucleotide composition of CLIP tags from three high *yw* ovary and *tud* embryo (0–2 h) Aub CLIP-seq libraries. piRNAs (23–29 nucleotides) are indicated by a dashed box. **g**, Genomic distribution of CLIP tags for three high *yw* ovary and *tud* embryo (0–2 h) Aub CLIP-seq libraries. Overlap of piRNAs from CLIP and immunoprecipitation libraries. All error bars denote s.d.; $n = 3$.

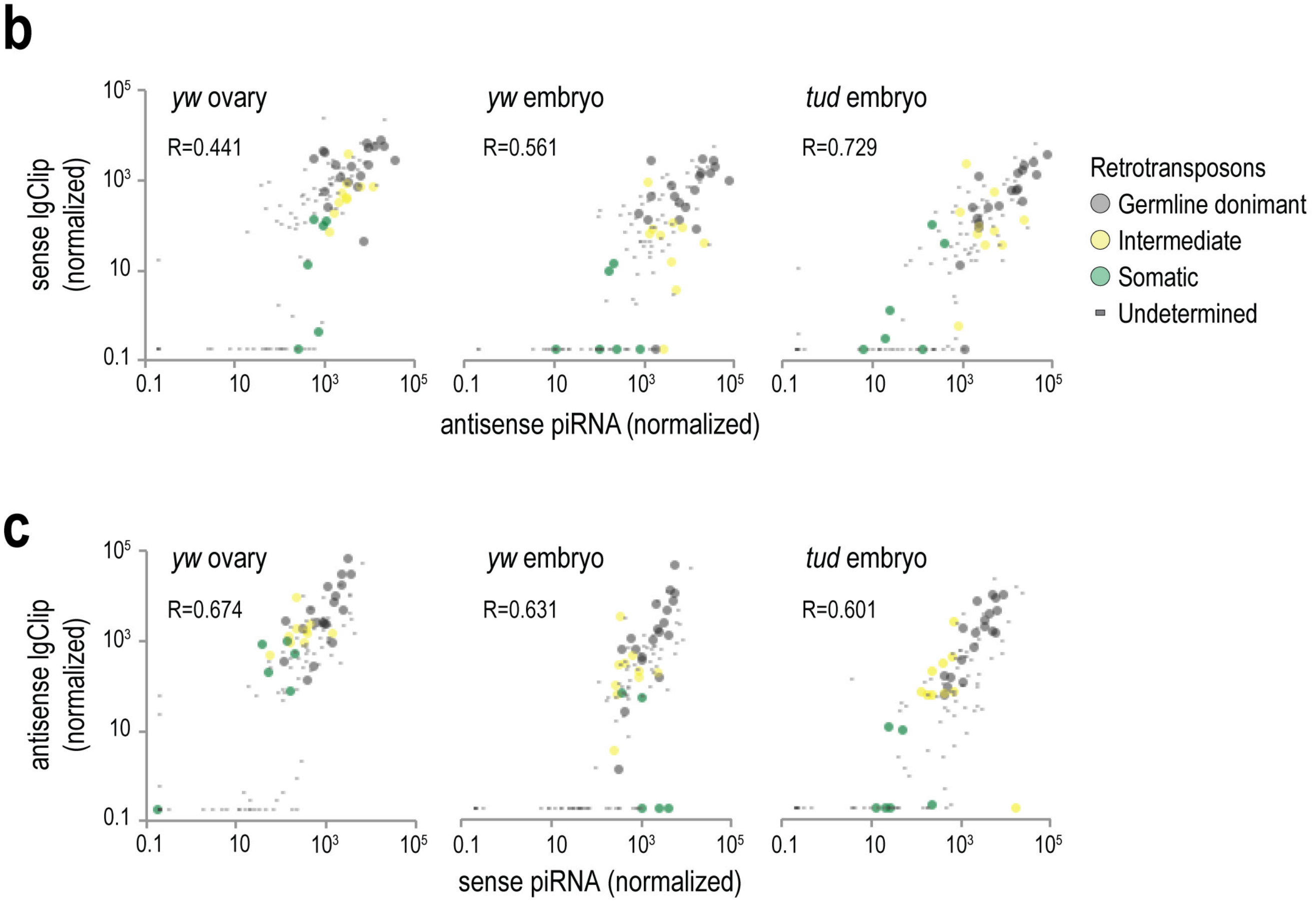


Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Pairwise comparisons of transposon piRNA populations from various libraries. **a**–**c**, Scatterplot comparison of normalized abundance of piRNAs mapped on consensus retrotransposon sequences (sense and antisense), from *yw* embryo (0–2 h) standard Aub immunoprecipitation and Aub CLIP libraries (**a**); from *yw* ovary libraries (**b**); and from *tud* embryo (0–2 h) libraries (**c**). Pearson correlation is shown for all elements in each plot. Retrotransposon categories are set as in ref. 31. **d**–**f**, Scatterplot comparison of normalized abundance of transposon-derived piRNAs in Aub CLIP libraries prepared from higher molecular mass signals (high; Fig. 1a, marked with a light blue line), with the piRNAs found in the libraries prepared from the main radioactive signal (low; Fig. 1a, marked with a dark blue line) from *yw* embryo (0–2 h) (**d**); from *yw* ovary Aub CLIP 'high' and 'low' libraries (**e**); and from *tud* embryo (0–2 h) Aub CLIP high and low libraries (**f**). These comparisons indicate that the piRNA loads in low and high CLIP libraries are essentially identical. **g**, Scatterplot comparison of normalized abundance of transposon-derived piRNAs for *yw* ovary and *tud* ovary Aub immunoprecipitation libraries, to evaluate changes of piRNA load in the absence of Tudor. While antisense-derived piRNAs are largely unchanged, a few sense-derived piRNAs are changed (blood retrotransposon is indicated). **h**, **i**, Scatterplot comparison of normalized abundance of transposon-derived piRNAs for *yw* ovary and *yw* embryo (0–2 h) Aub immunoprecipitation libraries (**h**); and for *tud* ovary and *tud* embryo (0–2 h) libraries (**i**).

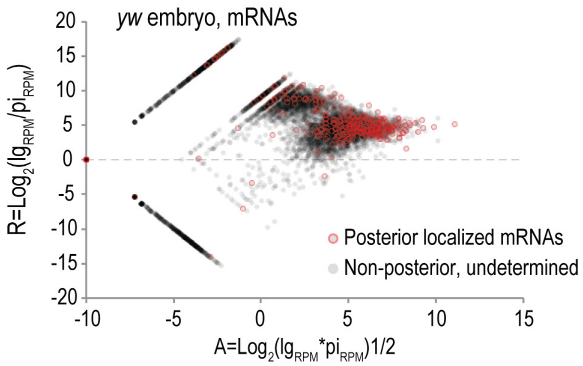
a

Library	total lgClips mapped within retrotransposons (consensus)	lgClips with overlapping antisense piRNAs	lgClips with overlapping antisense piRNAs (%)	average (%)
dme_Aub_CLIP_tud_embryo_0-2h_H1	3866282	785302	20.31155513	27.18524074
dme_Aub_CLIP_tud_embryo_0-2h_H2	3832702	1130647	29.49999765	
dme_Aub_CLIP_tud_embryo_0-2h_H3	3601118	1143145	31.74416945	
dme_Aub_CLIP_yw_embryo_0-2h_H1	886069	66353	7.488468731	15.62595067
dme_Aub_CLIP_yw_embryo_0-2h_H2	2558390	471642	18.43510958	
dme_Aub_CLIP_yw_embryo_0-2h_H3	1784247	373876	20.95427371	
dme_Aub_CLIP_yw_ovary_H1	858940	138301	16.10135749	21.03149503
dme_Aub_CLIP_yw_ovary_H2	999022	151961	15.21097633	
dme_Aub_CLIP_yw_ovary_H3	1231405	391367	31.78215128	



Extended Data Figure 3 | Retrotransposon targeting by complementary piRNAs identified by Aub CLIP. a, Overlap of lgCLIPs with complementary piRNAs from CLIP libraries, mapping on retrotransposons. **b**, **c**, Scatterplots of normalized abundance of antisense piRNAs and sense lgCLIPs (**b**) and for sense piRNAs and antisense lgCLIPs (**c**) mapped on retrotransposons for the indicated Aub CLIP libraries. Pearson correlation is shown for all elements in every plot. Retrotransposon categories are set as in ref. 31.

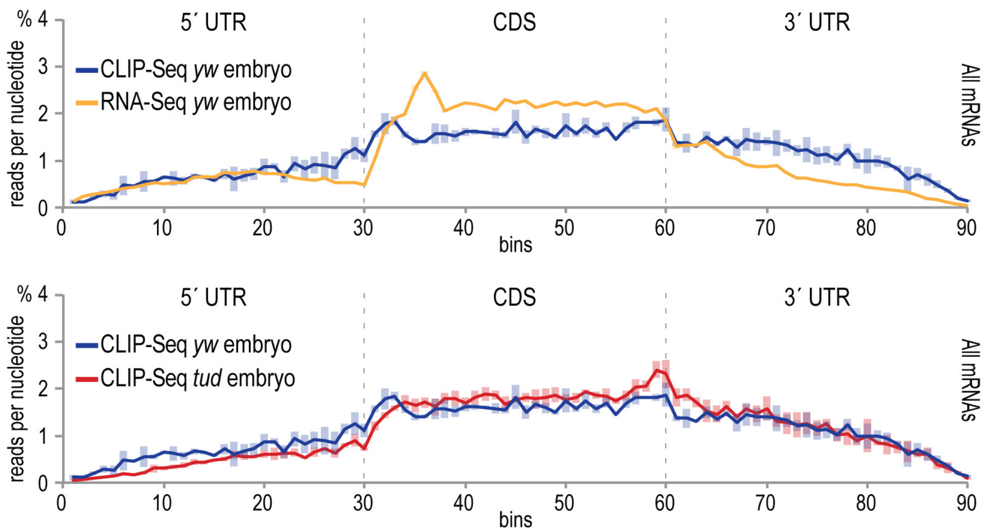
a



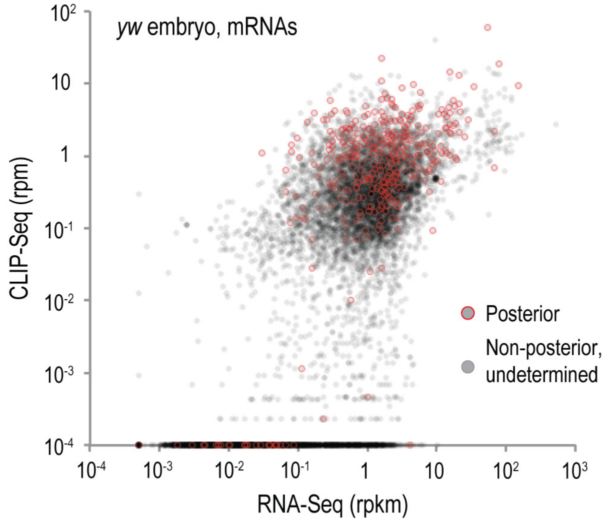
b

RNA-Seq library	Number of sequenced reads	Number of mapped reads
yw_embryo_0-2h_1	50,017,107	36,677,726
yw_embryo_0-2h_2	24,011,351	18,364,321
tud_embryo_0-2h_1	10,098,572	9,283,256
tud_embryo_0-2h_2	11,677,772	10,217,987
aub_embryo_0-2h_1	43,856,239	30,767,884
aub_embryo_0-2h_2	42,343,127	26,912,631

c

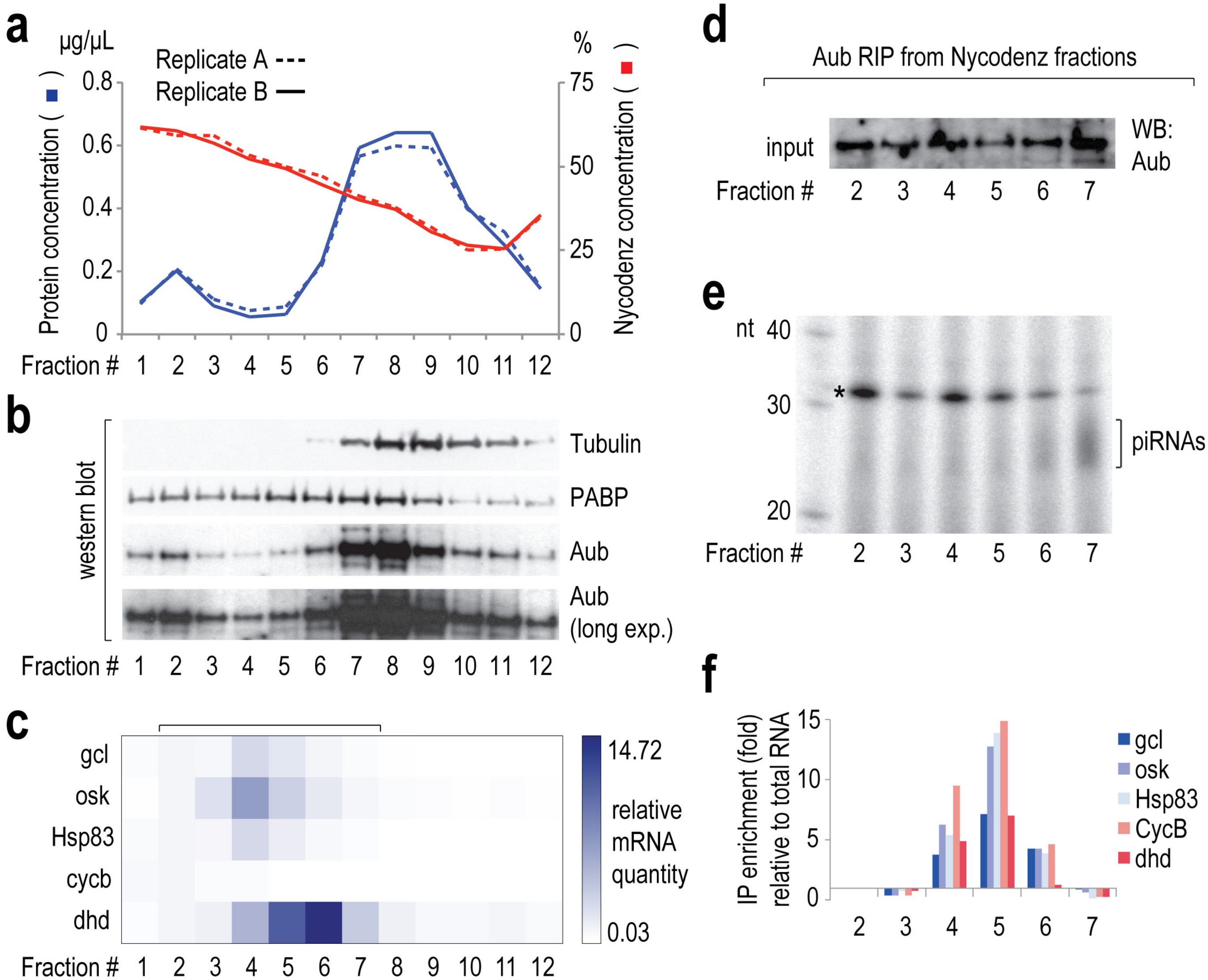


d



Extended Data Figure 4 | See next page for caption.

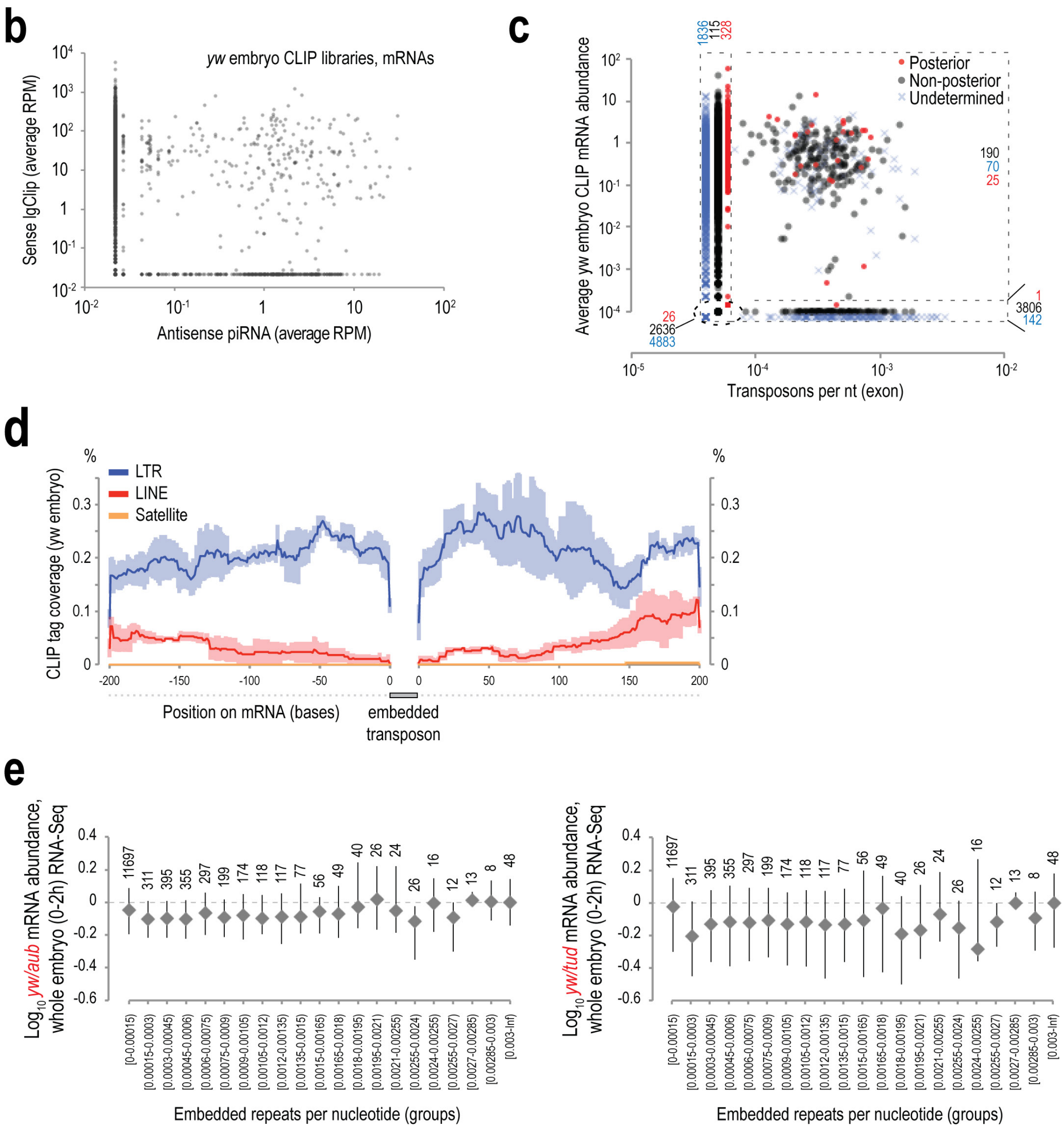
Extended Data Figure 4 | CLIP identifies extensive mRNA binding by Aub. a, Ratio average plot of normalized (reads per million, RPM) Aub CLIP tag (pi, piRNA; lg, lgCLIP) abundance (*A* value) versus lgCLIPs over piRNA abundance (*R* value), for all mRNAs. Outlined circles (red) correspond to genes that belong in the 12 posterior localization categories depleted in *tud* versus *yw* Aub CLIP libraries. Zero values are substituted with a small (smallest than the minimum) value so that log calculations are possible. This graph strongly suggests that mRNA binding by Aub as captured by CLIP is not for piRNA biogenesis purposes. **b**, Sequenced and mapped reads of RNA-seq libraries prepared in this study. **c**, Density of Aub CLIP-seq tags (*yw* embryo, and bottom panel: *tud* embryo) and RNA-seq reads (top panel: *yw* embryo) within the UTRs and coding sequences of the meta-mRNA. Each mRNA region is divided in 30 bins, and the number of the chimaeric mRNA fragments (genomic coordinate of the mRNA fragment midpoint) mapped within each bin is counted. Error bars indicate one s.d., $n = 3$ for CLIP-seq; minimum and maximum values for the two RNA-seq replicate libraries. **d**, Scatterplot of average normalized mRNA abundance for *yw* embryo RNA-seq (RPKM) and Aub CLIP-seq (RPM). Aub highly bound mRNAs with posterior localizations (Supplementary Table 4) are marked with a red circle. Zero values are substituted with a small (smallest than the minimum) value so that log calculations are possible. CLIP-seq identifies mRNAs that span the whole expression range of RNA-seq libraries, indicating that Aub CLIP does not capture transcripts simply based on abundance.



Extended Data Figure 5 | Partial purification of Aub RNPs from early embryo supports binding of germ plasm mRNAs by Aub. **a**, Fractionation of isopycnic Nycodenz density gradients of post-nuclear *yw* embryo lysate. Protein and Nycodenz concentration for every fraction is plotted. **b**, Western blot detection of indicated proteins in gradient fractions. A short and a long exposure (exp.) for Aub is shown. Uncropped gels for **b**, **d** and **e** can be found in Supplementary Fig. 1. **c**, Heat map of levels of indicated germ plasm mRNAs determined by quantitative RT–PCR (qRT–PCR), normalized to spiked luciferase RNA, and with fraction 2 as a reference. **d**, Western blot detection of Aub in indicated diluted Nycodenz fractions used for Aub RNA immunoprecipitation. **e**, Electrophoretic analysis on denaturing polyacrylamide gel of ^{32}P-labelled small RNAs immunoprecipitated with Aub from indicated gradient fractions. A bracket denotes piRNAs, detected primarily in fractions 6 and 7 (asterisk denotes 2S rRNA). **f**, Bar chart showing fold enrichment (over fraction-extracted total RNA) of indicated germ plasm mRNAs in Aub immunoprecipitations from gradient fractions, measured by qRT–PCR. Luciferase mRNA was used as a spike.

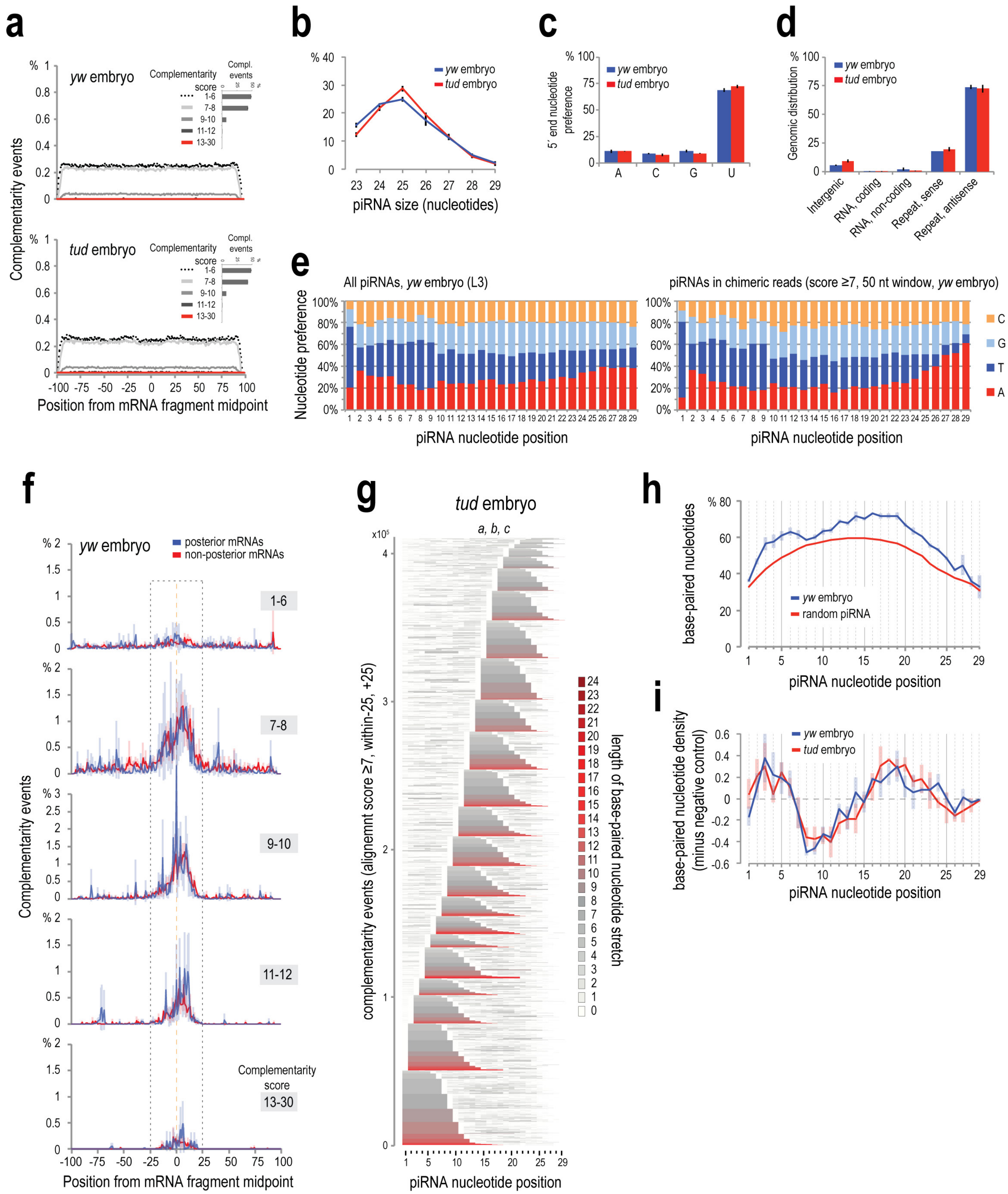
a

Library	total IgClips mapped within mRNAs	IgClips with overlapping antisense piRNAs	IgClips with overlapping antisense piRNAs (%)	average
dme_Aub_CLIP_tud_embryo_0-2h_H1	1000054	447	0.0%	0.1%
dme_Aub_CLIP_tud_embryo_0-2h_H2	909433	1300	0.1%	
dme_Aub_CLIP_tud_embryo_0-2h_H3	416111	941	0.2%	
dme_Aub_CLIP_yw_embryo_0-2h_H1	131081	141	0.1%	0.2%
dme_Aub_CLIP_yw_embryo_0-2h_H2	264642	369	0.1%	
dme_Aub_CLIP_yw_embryo_0-2h_H3	290583	1096	0.3%	
dme_Aub_CLIP_yw_ovary_H1	151394	281	0.2%	0.3%
dme_Aub_CLIP_yw_ovary_H2	148381	179	0.1%	
dme_Aub_CLIP_yw_ovary_H3	175321	1171	0.6%	



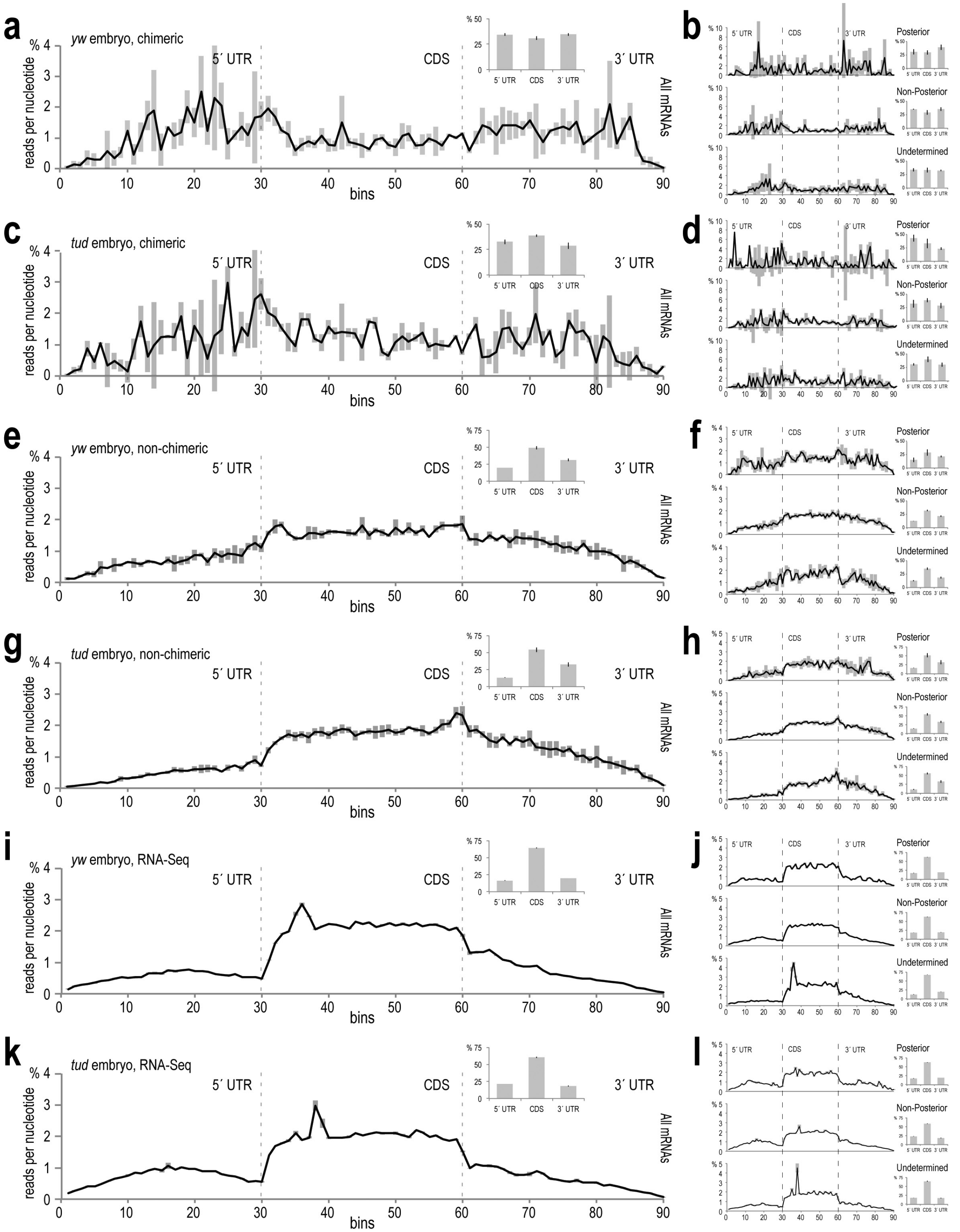
Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | Analysis of Aub CLIP tags mapping to mRNAs with regard to the presence of mRNA embedded transposons. a, Overlap of lgCLIPs with complementary piRNAs from CLIP libraries, mapping on mRNAs. **b**, Scatterplot of *yw* embryo Aub lgCLIPs mapped in the sense orientation on mRNAs, with piRNAs mapped in the antisense orientation. Zero values are substituted with a small (smallest than the minimum) value so that log calculations are possible. Contrary to retrotransposons (Extended Data Fig. 3), there is no correlation, suggesting that extensive piRNA complementarity cannot explain the widespread mRNA binding shown by mRNA lgCLIPs. **c**, Scatterplot of *yw* embryo Aub lgCLIPs mapped in the sense orientation on mRNAs with per base (nucleotide) mRNA embedded retrotransposons (LINE, long terminal repeat (LTR), satellite). Posterior, non-posterior and undetermined localizations are marked as indicated. The graph is separated into four quadrants: clockwise from bottom left corner: 0 embedded repeats, 0 CLIP tags; 0 embedded repeats, >0 CLIP tags; >0 embedded repeats, >0 CLIP tags, >0 embedded repeats, 0 CLIP tags. The number of genes in the four quadrants is indicated. Zero values are substituted with a small (smallest than the minimum) value (different small value for every localization category was used for clarity) so that log calculations are possible. This graph suggests that there is no correlation between the numbers of CLIP tags and embedded repeats within the mRNAs. **d**, Aub lgCLIPs density surrounding (±200 bases) mRNA-embedded retrotransposons (LINE, LTR, satellite as indicated). This analysis shows that there is no increase in the lgCLIP density in the areas flanking embedded repeats, suggesting that repeat sequences are not used as enriched target areas for mRNA binding by Aub. Error bars denote s.d.; $n = 3$. **e**, Analysis of mRNA expression level in relation to the number of embedded repeats. The number of embedded repeats per nucleotide of exon was plotted with the ratio ($\log_{10}$) of mRNA expression in *yw* embryo (0–2 h) versus $aub^{HN2/QC42}$ embryo (0–2 h) (left), and *yw* embryo (0–2 h) versus *tud* embryo (0–2 h) (right). The mRNAs are divided into groups based on the number of embedded repeats. The number above each data point denotes the number of mRNAs in each group. The graphs suggest that there is no proportional or consistent abundance change, decrease or increase, with the number of embedded repeats.



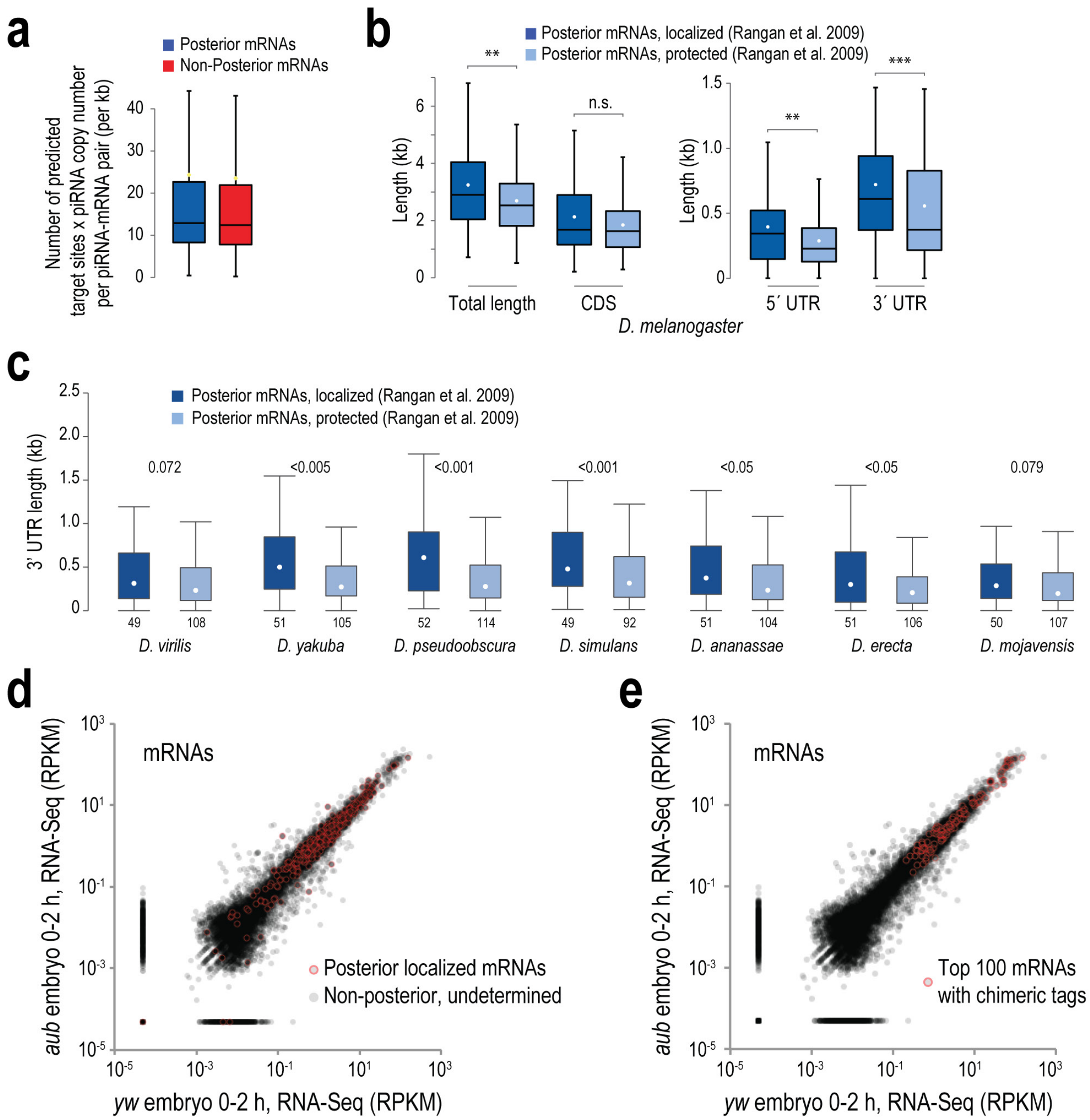
Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Characteristics of piRNAs and piRNA base-pairing with complementary target sites identified from analysis of chimaeric CLIP tags. a, piRNA–mRNA complementarity events for a random piRNA (negative control, average of three *yw* (top) and *tud* (bottom) embryo (0–2 h) samples), within ±100 bases from the midpoint of the mRNA part of the chimaeric read. Complementarity events are plotted per alignment score group as indicated, for clarity. Inset (per sample): bar chart of average complementarity events per score group. **b**, Size distribution of the piRNAs identified within chimaeric CLIP tags, for *yw* and *tud* embryo CLIP libraries. Only the piRNAs implicated in the complementarity events occurring within ±25 nucleotides from the midpoint of the mRNA fragment and with score ≥7 are analysed in this graph, and the graphs in **c**–**e**, **g**–**i**. **c**, 5′ end nucleotide preference for the piRNAs identified within chimaeric CLIP tags, for *yw* and *tud* embryo Aub CLIP libraries. **d**, Genomic distribution for the piRNAs identified within chimaeric CLIP tags, for *yw* and *tud* embryo Aub CLIP libraries. **e**, Per position nucleotide preference for all piRNAs in Aub *yw* embryo (0–2 h) CLIP library L3 (left), and for the piRNAs identified within chimaeric CLIP tags, for *yw* and *tud* embryo Aub CLIP libraries. **f**, Complementarity events between piRNAs and mRNA fragments of chimaeric reads, for posterior and non-posterior localized mRNAs (*yw* embryo). The plots are separated per score group. **g**, Heat maps showing base-paired nucleotides of piRNAs for all complementarity events identified within chimaeric CLIP tags (events occurring within ±25 nucleotides from mRNA fragment midpoint, score ≥7) for *tud* embryo. Colour is according to the length of the consecutive stretch of base-paired nucleotides that runs over every position (colour code shown on the right). Stacked piRNAs are aligned at their 5′ ends and sorted (bottom to top) following these rules: (a) starting position of the longest stretch of consecutive base paired nucleotides, relative to the piRNA end; (b) length of longest base-paired stretch; (c) total number of base-paired nucleotides. **h**, Base-pairing frequency along the piRNA length for *yw* embryo libraries (blue) and their negative control (red). **i**, Net base-pairing frequency along the piRNA length (red) and net density of base paired nucleotides (grey) in mRNAs from chimaeric CLIP tags from *tud* embryo libraries. All error bars denote s.d.; $n = 3$.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Non-chimaeric Aub CLIP tag (lgCLIP), chimaeric mRNA fragment and RNA-seq read density along the untranslated and coding sequences of mRNAs. a, Average density of chimaeric mRNA fragments (Aub CLIP, *yw* 0–2-h embryo) along the three parts of the meta-mRNA. Each mRNA region is divided in 30 bins and the number of the chimaeric mRNA fragments (genomic coordinate of the mRNA fragment midpoint) mapped within each bin is counted. Inset: bar plot showing cumulative density in each mRNA region. **b**, Average density of the chimaeric mRNA fragments on mRNA regions; mRNAs are separated into three localization groups as indicated: posterior localized (12 categories; Supplementary Table 3), non-posterior and undetermined localization. Inset: bar plot showing cumulative density in each mRNA region. **c**, As in **a** for chimaeric mRNA fragments from Aub CLIP libraries, *tud* embryo (0–2 h). **d**, As in **b** for chimaeric mRNA fragments from Aub CLIP libraries, *tud* embryo (0–2 h). **e**, As in **a** for non-chimaeric lgCLIPs from Aub CLIP libraries, *yw* embryo (0–2 h). **f**, As in **b** for non-chimaeric lgCLIPs from Aub CLIP libraries, *yw* embryo (0–2 h). **g**, As in **a** for non-chimaeric lgCLIPs from Aub CLIP libraries, *tud* embryo (0–2 h). **h**, As in **b** for non-chimaeric lgCLIPs from Aub CLIP libraries, *tud* embryo (0–2 h). **i**, As in **a** for RNA-seq reads, *yw* embryo (0–2 h). **j**, As in **b** for RNA-seq reads, *yw* embryo (0–2 h). **k**, As in **a** for RNA-seq reads, *tud* embryo (0–2 h). **l**, As in **b** for RNA-seq reads, *tud* embryo (0–2 h). Error bars denote s.d.; $n = 3$.

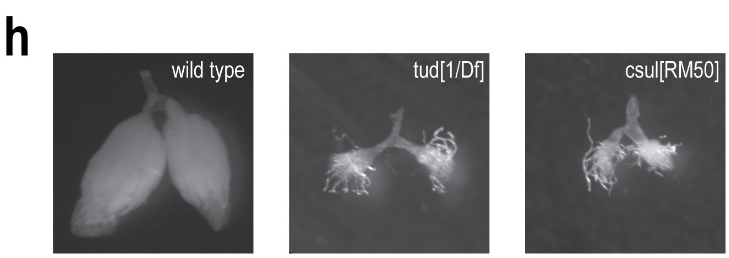


f

Hatch Rate			
Maternal Phenotype	Total Eggs	Hatched	Percentage
wild type	385	340	88.31
aub[HN2/QC42]	316	0	0.00
tud[1/Df]	297	156	52.53
csul[RM50]	384	164	42.71

g

Progeny Fertility			
Maternal Phenotype	Total Flies	Fertile	Percentage
wild type	50	50	100.00
aub[HN2/QC42]	0	0	0.00
tud[1/Df]	139	0	0.00
csul[RM50]	50	0	0.00



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Lengths of posterior localized mRNAs in *Drosophila* species; characteristics of embryos used in our studies. **a**, Box-and-whisker plot of the number of predicted piRNA target sites (per kilobase of mRNA sequence) for every mRNA–piRNA pair, multiplied by the piRNA copy number. Posterior and non-posterior mRNAs are as indicated. Black lines denote the median. This graph indicates that the 'targeting potential' (number of predicted complementary sites multiplied by the piRNA copy number) of each piRNA against each mRNA is the same for the two localization categories, suggesting that the piRNA copy number is not a contributing factor for the observed preference of posterior localized mRNAs for piRNA adhesion. **b**, Box-and-whisker plot of the lengths of *D. melanogaster* mRNAs (and their 5′ UTR, coding sequences and 3′ UTR parts) that are found in the enriched and protected categories, as defined previously[10]. Black lines denote the median; white dots denote the mean. n.s., not significant ($P > 0.05$); $^{**}P < 0.01$; $^{***}P < 0.001$; one-sided Wilcoxon rank sum test. **c**, Box-and-whisker plot of the lengths of the 3′ UTRs of mRNAs from the indicated *Drosophila* species that are orthologous to the *D. melanogaster* mRNAs found in the localized and protected categories, as defined previously[10]. Incomplete annotation did not allow us to perform this analysis for all the species shown in Fig. 4i.White dots denote the mean. *P* values of the statistical test (one-sided Wilcoxon test) of whether the lengths of the localized versus protected mRNAs are different, are shown for each species. **d**, **e**, RNA-seq scatterplots from 0–2-h wild-type (*yw*) and 0–2-h Aub-null (*aub*) embryos. Shown in red are posterior localized mRNAs (**d**) or the top 100 mRNAs identified from Aub CLIP piRNA–mRNA chimaeric reads (**e**). There is no change in mRNA levels between wild-type and *aub* mutant 0–2-h embryos. **f**, **g**, Hatch rates (**f**) and fertility of progeny (**g**) of embryos from indicated genotypes. Note that, unlike Tudor and Csul, the absence of Aub ($aub^{HN2/QC42}$) leads to complete embryo lethality. **h**, Gross ovary appearance of wild-type (*yw*), Tudor mutant (*tud[1/Df])* and Csul mutant ($csul^{RM50}$) adult flies. Note complete absence of germline ovarian tissue in adult flies lacking Tudor or Csul; embryos from these flies develop into agametic adults because PGCs are never specified.

Extended Data Table 1 | Overlap of piRNAs from CLIP and immunoprecipitation libraries

Library 1	Library 2	unique piRNA sequences in library 1	unique piRNA sequences in library 2	common	percent1	percent2	average percent 2
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_H1	6913438	348812	150654	2.179147336	43.19060124	42.22806
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_H2	6913438	838891	333876	4.829377222	39.79968792	
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_H3	6913438	694458	284532	4.115636822	40.97180823	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_H1	9938639	560082	286627	2.883966306	51.17589924	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_H3	9938639	293375	156976	1.579451673	53.50694504	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_H2	9938639	332484	176012	1.770986953	52.93848727	
Aub_IP_tud_embryo_0-2h	Aub_CLIP_tud_embryo_0-2h_L1	5147948	1257672	458182	8.900284152	36.43096133	
Aub_IP_tud_embryo_0-2h	Aub_CLIP_tud_embryo_0-2h_H2	5147948	1104187	460392	8.943213879	41.69511143	
Aub_IP_tud_embryo_0-2h	Aub_CLIP_tud_embryo_0-2h_H3	5147948	2567880	948630	18.42734231	36.94214683	
Aub_IP_tud_embryo_0-2h	Aub_CLIP_tud_embryo_0-2h_H1	5147948	1040626	379030	7.362739484	36.4232683	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_L1	9938639	1850192	874693	8.800933407	47.27579624	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_L2	9938639	2407082	1108175	11.15016855	46.03810755	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_L3	9938639	3082922	1367516	13.75959022	44.35778784	
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_L1	6913438	2012094	722743	10.45417634	35.91994211	
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_L2	6913438	2161685	769241	11.12675054	35.58524947	
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_L3	6913438	2701578	902250	13.0506703	33.39714789	

Comparisons of piRNA sequences found in CLIP and immunoprecipitation libraries from same tissues.

doi:10.1038/nature17163

Crystal structure of a substrate-engaged SecY protein-translocation channel

Long Li[1]*, Eunyong Park[1]†*, JingJing Ling[2], Jessica Ingram[2], Hidde Ploegh[2] & Tom A. Rapoport[1]

Hydrophobic signal sequences target secretory polypeptides to a protein-conducting channel formed by a heterotrimeric membrane protein complex, the prokaryotic SecY or eukaryotic Sec61 complex. How signal sequences are recognized is poorly understood, particularly because they are diverse in sequence and length. Structures of the inactive channel show that the largest subunit, SecY or Sec61α, consists of two halves that form an hourglass-shaped pore with a constriction in the middle of the membrane and a lateral gate that faces lipid[1–10]. The cytoplasmic funnel is empty, while the extracellular funnel is filled with a plug domain. In bacteria, the SecY channel associates with the translating ribosome in co-translational translocation, and with the SecA ATPase in post-translational translocation[11]. How a translocating polypeptide inserts into the channel is uncertain, as cryo-electron microscopy structures of the active channel have a relatively low resolution (∼10 Å) or are of insufficient quality[6–8]. Here we report a crystal structure of the active channel, assembled from SecY complex, the SecA ATPase, and a segment of a secretory protein fused into SecA. The translocating protein segment inserts into the channel as a loop, displacing the plug domain. The hydrophobic core of the signal sequence forms a helix that sits in a groove outside the lateral gate, while the following polypeptide segment intercalates into the gate. The carboxy (C)-terminal section of the polypeptide loop is located in the channel, surrounded by residues of the pore ring. Thus, during translocation, the hydrophobic segments of signal sequences, and probably bilayer-spanning domains of nascent membrane proteins, exit the lateral gate and dock at a specific site that faces the lipid phase.

To determine the structure of an active SecY channel, we initially generated in *Escherichia coli* a translocation intermediate, consisting of SecA, SecY complex, and a short segment of a secretory protein fused to a fast-folding green fluorescent protein (GFP) (Extended Data Fig. 1a). Although this complex could be purified[12], it failed to crystallize. We therefore reduced the complexity of the system by fusing a short segment of a secretory protein directly into SecA. The segment contains the signal sequence of OmpA and a short polypeptide following it, and was inserted into the tip of the two-helix finger of SecA (SecA-OAIns; Fig. 1a and Extended Data Fig. 1b), because the finger was seen to protrude into the cytoplasmic cavity of SecY in a structure of SecA/SecY complex lacking a translocation substrate[9]. Using *E. coli* SecA-OAIns and *E. coli* SecY complex, the inserted secretory protein segment was indeed translocated to the periplasm in *E. coli*, as demonstrated by the formation of a disulfide bridge between a cysteine introduced C-terminally of the signal sequence and a cysteine placed into the plug domain of SecY (Extended Data Fig. 2a). This disulfide bridge formed spontaneously; that is, without addition of an exogenous oxidant. The introduction of Gln residues into the hydrophobic core of the signal sequence abolished disulfide bridge formation (Extended Data Fig. 2b), demonstrating that an intact signal sequence is required for translocation of the polypeptide segment. Similar results were obtained with *Bacillus subtilis* SecA-OAIns and *Geobacillus thermodenitrificans* SecYE (Extended Data Fig. 2c), a complex of increased thermostability that is functional in *E. coli*. After optimization (Extended Data

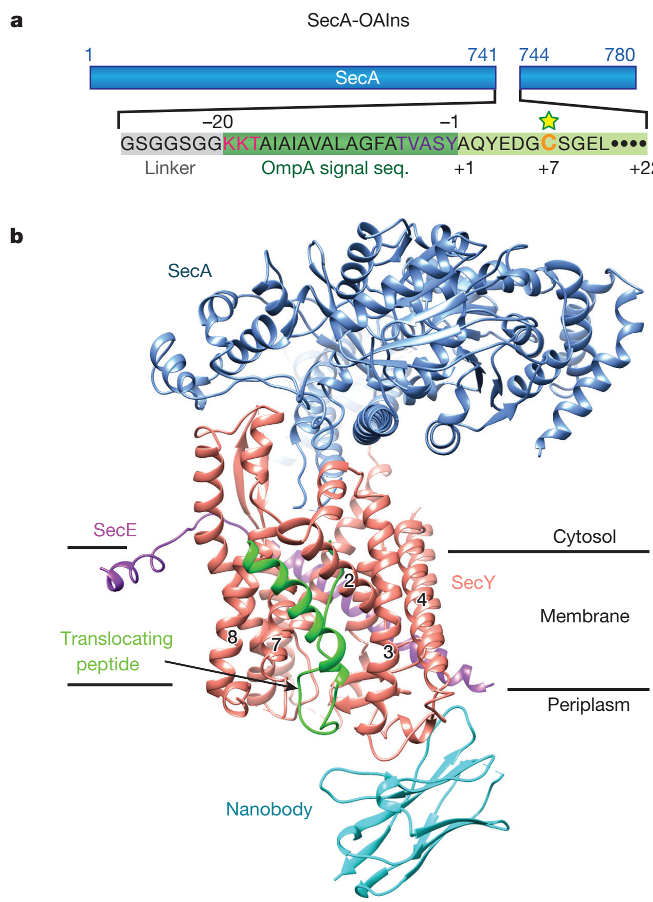


Figure 1 | Architecture of the active SecY channel. a, A secretory protein segment was inserted into the two-helix finger of the SecA ATPase (SecA-OAIns). The segment contains a linker (grey), the signal sequence of OmpA, consisting of the N-, H-, and C-regions (in red, black, and purple letters, respectively), and a region (in light green) that includes a unique cysteine (yellow star). Residues in the signal sequence are numbered backwards from the cleavage site. The fused segment inserts into the SecY channel *in vivo* and spontaneously forms a disulfide bridge with a cysteine in the plug. This complex was used for structure determination. **b**, Ribbon diagram of the complex, viewed from the side. The numbers refer to TMs of SecY. The lines indicate the membrane boundaries. A nanobody was used for crystallization.

[1]Howard Hughes Medical Institute and Harvard Medical School, Department of Cell Biology, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. †Present address: The Rockefeller University and Howard Hughes Medical Institute, 1230 York Avenue, New York, New York 10065, USA.
*These authors contributed equally to this work.

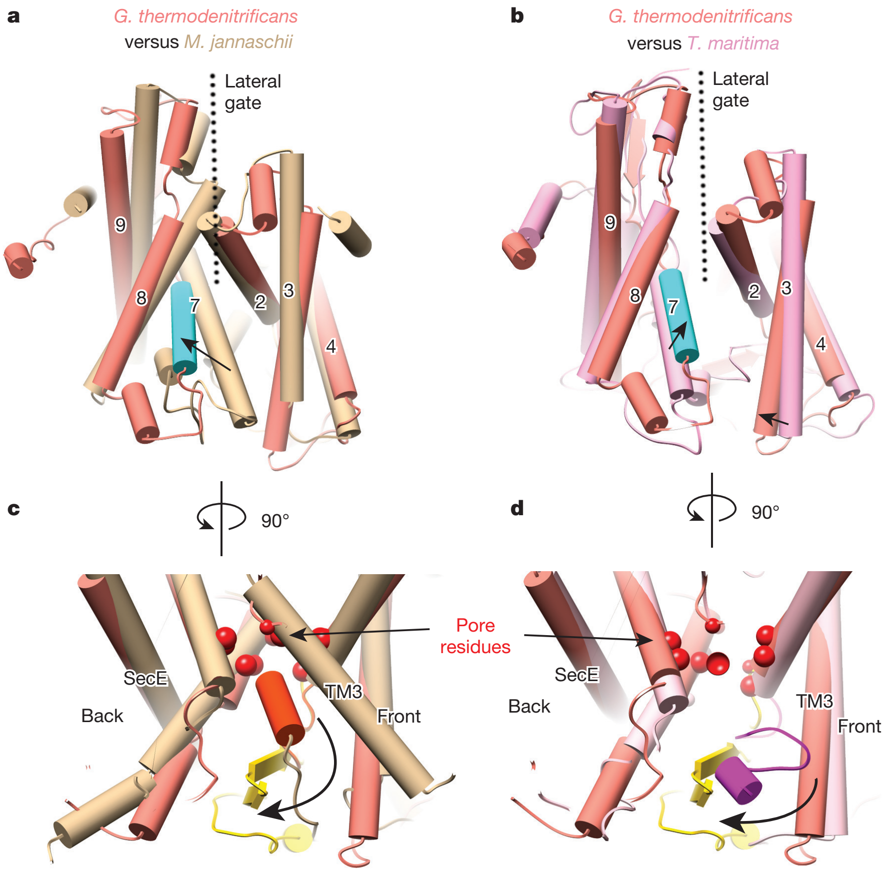


Figure 2 | Conformational changes of the SecY channel. a, Comparison of the lateral gate of the active *G. thermodenitrificans* channel (salmon, except for TM7 in cyan) with the closed SecY channel of *M. jannaschii* (in tan). The largest changes are indicated by arrows. The translocating peptide segment is omitted for clarity. **b**, As in **a**, but a comparison with that of the SecA-bound *T. maritima* channel lacking a translocating polypeptide (pink). **c**, Comparison of the plugs (orange for idle *M. jannaschii* channel and yellow for the active *G. thermodenitrificans* channel). Pore residues are shown as red spheres. **d**, As in **c**, but comparison with the inactive *T. maritima* channel (plug in magenta).

Fig. 2d–f), the construct chosen for crystallization contained 49 residues inserted into the two-helix finger of *B. subtilis* SecA, with a cysteine at position +7 in the region following the signal sequence of 20 residues. Channel insertion of the secretory protein segment was similar to that observed with the physiological system, containing wild-type SecA and a GFP fusion to a secretory protein fragment (Extended Data Fig. 1), except that the latter required an additional polypeptide segment to span the SecA molecule. Thus, our simplified system is a faithful mimic of normal initiation of protein translocation. Binding of SecA to the SecY complex seems to be sufficient to cause polypeptide chain insertion into the channel, similar to how ribosome binding allows nascent chain insertion in co-translational translocation[13]. In our system, disulfide crosslinking at the periplasmic side made channel insertion irreversible.

The disulfide-bridged complex of *B. subtilis* SecA-OAIns and *G. thermodenitrificans* SecYE was purified and crystallized in the presence of ADP and BeFx (Extended Data Fig. 3), conditions that lock SecA into a conformation close to its ATP-bound state and maximize the affinity of SecA for the channel[9,14]. The diffraction of the crystals was improved by the use of single-domain antibody fragments (nanobodies), raised against *G. thermodenitrificans* SecYE and selected for binding to periplasmic loops of the SecA-OAIns/SecYE complex, and by soaking crystals with a Ta_6Br_{12} metal ion cluster. The structure was determined from multi-wavelength anomalous diffraction (MAD) data obtained with a crystal that diffracted to a resolution of 3.70 Å along one axis and 4.48 Å along the other two (Fig. 1b, Extended Data Fig. 4 and Extended Data Table 1). An initial experimental electron density map had a resolution of ~5.5 Å. This map was improved by density modification and molecular replacement using higher-resolution structures of SecA, SecYE, and the nanobody, followed by cycles of model building and refinement. Inclusion of the model-refined Ta_6Br_{12} clusters as the resolved heavy atom substructure for recalculation of MAD phases did not further improve the map. Nevertheless, the final map allowed the unambiguous placement of all transmembrane segments (TMs) of SecY and SecE and of many other regions. The translocating polypeptide segment could be built into the map without model bias (Extended Data Fig. 4c). As expected, the nanobody bound to the periplasmic side of SecY, interacting with both the plug and the loop between TM3 and TM4 (Fig. 1b and Extended Data Fig. 5a).

The structure of the active channel shows that SecA undergoes relatively small conformational changes compared with a *Thermotoga maritima* complex lacking a translocating chain[9] (Extended Data Fig. 6). SecA binds to the cytoplasmic loop between TM8 and TM9 and the C-terminal tail of SecY (Extended Data Fig. 5b, c). It probably binds only weakly to the loop between TM6 and TM7, as its tip is disordered. In contrast to the ribosome[6–8], SecA also binds to the amino (N)-terminal half of SecY; that is, the loop between TM2 and TM3 (Extended Data Fig. 5b). Thus, at least in its ATP-bound state, SecA prevents large relative movements of the two halves of SecY.

SecY also undergoes relatively small changes, except at the lateral gate (Fig. 2). Compared with the idle *Methanocaldococcus jannaschii* or *Thermus thermophilus* channels[1,3], only TM7 (*M. jannaschii*) or TM7 and TM8 (*T. thermophilus*) significantly shift their positions (Fig. 2a and Extended Data Fig. 7a). Compared with the SecA-bound *T. maritima* channel[9], the periplasmic ends of TM3 and TM7 move towards each other, and TM7 tilts by 10° relative to the plane of the membrane (Fig. 2b), changes that generate a pocket for the signal sequence (see below). In both structures, the lateral gate is partly open (compare Fig. 2a and Fig. 2b).

In the active *G. thermodenitrificans* SecY channel, the plug consists of two β-strands and therefore differs from the α-helical structures observed in other species[1,4,5,9] (Extended Data Fig. 8a). Such variability is consistent with the fact that the amino-acid sequence of the plug region is least conserved[1], and that plug deletions cause neighbouring polypeptide regions to form new plug domains[15]. Different plug structures can probably be tolerated, as long as they fill the extracellular cavity of the channel, so that the closed state of the channel is stabilized and small molecules cannot pass through it.

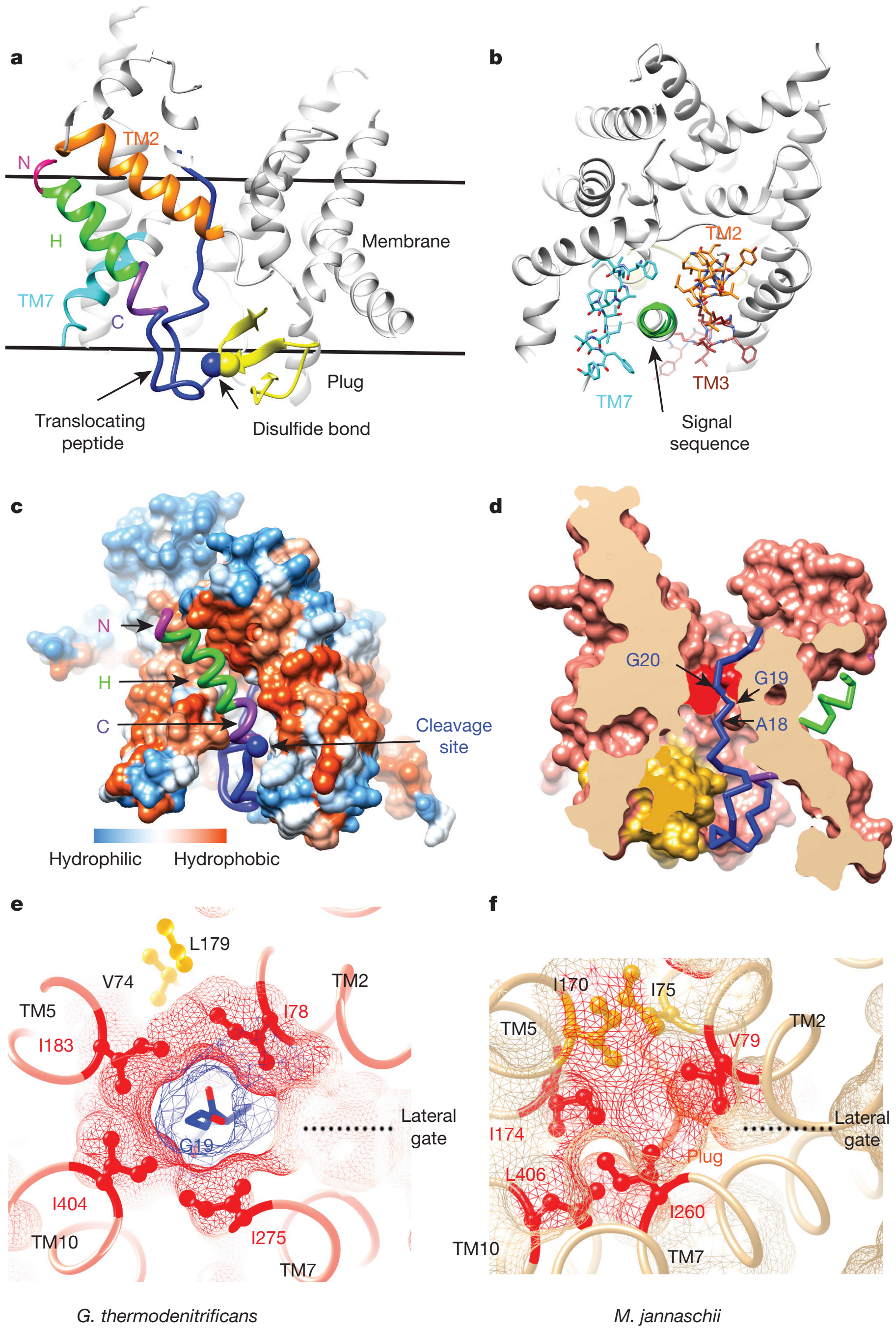


Figure 3 | The polypeptide in the channel. **a**, Cut-away side view of the translocating polypeptide. The N-, H-, and C-regions of the signal sequence are indicated. **b**, View perpendicular to the axis of the signal sequence helix. **c**, Side view showing the translocating polypeptide together with a surface representation of the channel, with hydrophilic and hydrophobic residues in blue and orange, respectively. **d**, Side view of a cut through a surface representation model, with the main chain of the translocating chain shown as blue sticks. The residues located in the pore ring (red) are indicated. The plug is in orange. **e**, Top view of a slab, showing pore residues and Gly19 of the translocating chain as stick and balls with a mesh surface representation. Pore residues displaced in the active channel are in yellow. **f**, As in **e**, but with the closed *M. jannaschii* channel. I75 and I170 correspond to V74 and L179 in *G. thermodenitrificans*, respectively.

However, it is possible that the plug has different conformations in the closed and active channels.

Whereas the plug is close to the central constriction in the closed *M. jannaschii* channel[1], in the active *G. thermodenitrificans* channel it moves to the periplasmic side and towards the back of the channel, away from the lateral gate (Fig. 2c). The plug comes close to the TM of SecE, consistent with disulfide crosslinking experiments[16,17]. In a SecA–SecY structure lacking a translocation substrate[9], the plug moves a smaller distance and towards the front (Fig. 2d), partly sealing the opened lateral gate (Extended Data Fig. 8b). In an intact membrane, this would prevent surrounding lipid molecules from moving through the lateral gate into the extracellular cavity. The plug is probably flexible in the active channel, but in our crystal structure it is confined both by the disulfide bond to the translocating chain and by the interaction with the nanobody (Extended Data Fig. 5a). Indeed, in a 6.5 Å-resolution structure determined without nanobody, the plug is shifted further towards the back, although there are otherwise only small differences (Extended Data Fig. 9a).

The signal sequence of the secretory protein segment forms a helix that is tilted ∼45° relative to the plane of the membrane (Fig. 3a). The positively charged N terminus of the signal sequence (N-region) is in the same plane as hydrophilic residues of SecY. In an intact membrane, the N-region could interact with the negatively charged head groups of the phospholipid bilayer. This interaction may retain the N terminus on the cytoplasmic side of the membrane while the C-terminal end of the signal sequence moves through the channel, resulting in loop insertion of the translocating polypeptide.

The hydrophobic core of the signal sequence (H-region; residues −17 to −6 of the original OmpA sequence; Fig. 1a) sits in a groove outside the lateral gate of SecY and forms a helix that runs almost parallel to TM2 (Fig. 3a, b). Some residues make van der Waals contacts with hydrophobic amino acids in TM2 (Extended Data Fig. 7c), but most

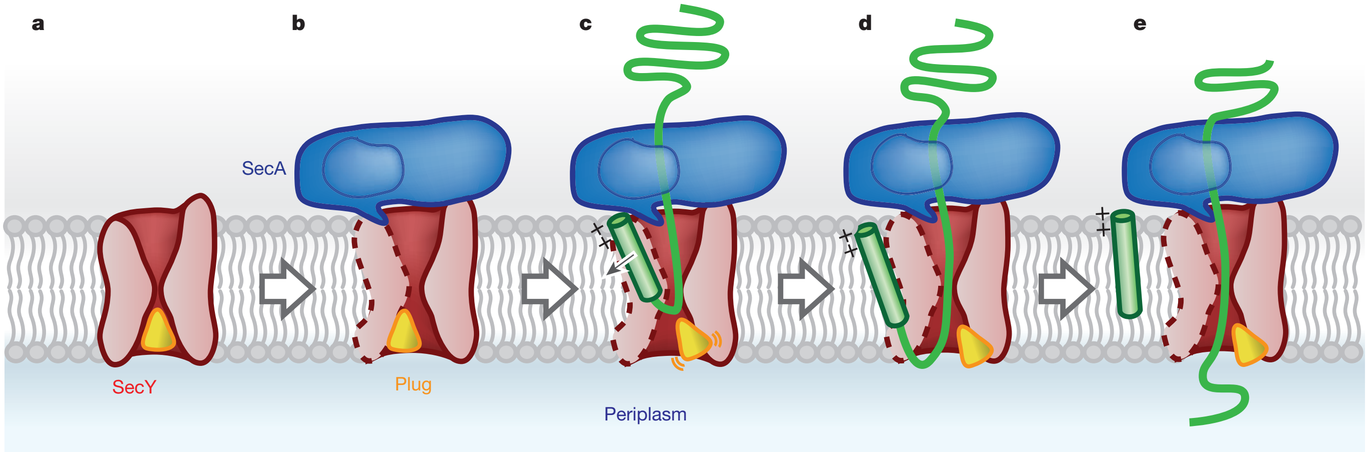


Figure 4 | Scheme of SecA-mediated protein translocation. Stage **a** corresponds to the closed channel[1–3], stage **b** to the structure of the inactive SecA–SecY complex[9], stage **c** represents an intermediate to stage **d**, which corresponds to the structure of the active channel. Stage **e** is attained after signal sequence cleavage. The translocating polypeptide is shown in green and the signal sequence as a green cylinder. The lateral gate of the channel is shown as a broken line on the left. The clamp of SecA is indicated.

face detergent (Fig. 3b), and would be in contact with hydrocarbon chains of phospholipids in an intact membrane.

The C-terminal region of the signal sequence (C-region; residues −5 to −1; Fig. 1a) replaces the periplasmic end of TM7 in the closed *M. jannaschii* channel (Extended Data Fig. 7b). This segment forms a distorted, amphipathic α-helix that is intercalated between TM7 and TM3 into the periplasmic side of the lateral gate (Fig. 3a, c). In our model, the side chains of Thr(−5) and Ser(−2) point into the periplasmic cavity previously occupied by the plug in the closed channel (Extended Data Fig. 7d). The hydrophobic residues Val(−4) and Ala(−3) face detergent/lipid. The lateral gate is thus sealed by the C-region from surrounding lipid molecules, which could otherwise pass through a large opening generated by the displacement of the plug from the front (Fig. 3c). After signal sequence cleavage, the periplasmic parts of TM7 and TM3 probably move towards each other and seal the lateral gate.

The hydrophilic polypeptide segment following the signal sequence adopts a partly extended conformation with a loop in the periplasmic cavity, centred on the cysteine used for crosslinking to the plug. The signal sequence cleavage site is located inside the channel (Fig. 3c), probably inaccessible to the periplasmically disposed active site of signal peptidase[18]. Thus, at some point during translocation, the C-region of the signal sequence probably must adopt a more extended conformation, a change also suggested by experiments with synthetic signal peptides[19,20].

The polypeptide chain inside the channel is perpendicular to the plane of the membrane (Fig. 3a). The two strands of the hairpin formed by the translocating polypeptide do not interact with one another, so that during translocation the C-terminal part of the hairpin could move unimpeded through the centre of the channel. Our model places residue Gly(+19), or one of the neighbouring residues (Ala18 or Gly20), of the translocating polypeptide inside the pore ring (Fig. 3d). The density around Gly19 is particularly strong, indicating that this segment is confined by the surrounding four pore ring residues (Ile78, Ile183, Ile275, Ile404). The ring is wider than in the idle *M. jannaschii* or SecA-bound *T. maritima* channels (diameters 8.8 Å versus 5.6 Å or 6.6 Å respectively). Crystallization may have favoured the presence of small amino acids in the pore, minimizing its expansion by the presence of a translocating chain. However, even a small increase in pore diameter would allow the passage of amino acids with larger side chains. The pore ring residues fit snugly around the translocating polypeptide (Fig. 3e), confirming that they form a 'gasket' that maintains the permeability barrier for ions and other small molecules during translocation[21]. Consistent with disulfide crosslinking experiments[22], only pore ring residues contact the translocating chain (Fig. 3d). Thus, the hourglass-shape of the channel minimizes interactions with the translocation substrate, facilitating its movement through the channel.

The pore ring of the idle *M. jannaschii* channel contains two additional residues (Ile75 and Ile170; Fig. 3f). In the active channel, the corresponding residues (Val74, Leu179) are displaced (Fig. 3e). The pore 'ring' is thus open at the lateral gate between Ile78 in TM2 and Ile275 in TM7 (Fig. 3e). These features suggest that a translocating polypeptide segment continuously encounters the hydrocarbon chains of surrounding lipids; when sufficiently hydrophobic, the segment will partition into the lipid phase and become a TM domain of a membrane protein[23,24].

Our crystal structure probably reflects the physiological situation of a translocating polypeptide. Five of the seven residues of the N-terminal linker are invisible and thus probably flexible, allowing unrestricted interaction of the signal sequence with the channel. In addition, most polypeptide segments following the signal sequence are in a relaxed conformation, unconstrained by fusion to SecA, or the disulfide bridge to the plug. The disulfide bridge helps to stabilize the signal sequence in the channel, but it probably does not lead to gross distortions, because the plug is mobile and the disulfide bridge is formed spontaneously *in vivo*.

The crystal structure of the active channel leads to a refined model for post-translational protein translocation in bacteria (Fig. 4). The SecY channel is initially in the idle state, with the plug in the centre and the lateral gate closed (Fig. 4a). Binding of SecA primes the channel for the arrival of a secretory protein precursor: the lateral gate is partly opened, the pore ring widened, and the plug domain moved towards the front (Fig. 4b). Next, the secretory protein inserts into the channel as a loop, with the C-terminal section of the polypeptide hairpin in the pore proper, surrounded by pore ring residues (Fig. 4c, d). During subsequent cycles of ATP hydrolysis, SecA uses a 'push-and-slide' mechanism to move the C-terminal part of the polypeptide loop through the pore[14]. Eventually, the signal sequence is cleaved by signal peptidase (Fig. 4e).

During loop insertion, the H-region of the signal sequence moves through the cytoplasmic part of the lateral gate and ends up in a hydrophobic groove on the outside, while the following hydrophilic segment crosses the lateral opening of the pore ring. A signal sequence might move through a partly open gate in an extended conformation, or it could move through a widened gate as a preformed helix. The latter possibility is suggested by a ~10 Å-resolution electron microscopy (EM) structure of a ribosome/nascent chain/channel complex, in which a signal sequence helix was seen inside the lateral gate[6], probably prevented from exit by a disulfide bridge between a cysteine at the end of the signal sequence and a cysteine in the plug. The groove on

the outside of the lateral gate appears to be a general binding site for hydrophobic sequences, as indicated by a 8.5 Å resolution structure of a complex in which the OmpA signal sequence was replaced by that of DsbA (Extended Data Fig. 9b, c), and EM structures of ribosome–channel complexes, in which TMs of nascent membrane proteins are located at about the same position[7,25] (Extended Data Fig. 7e). Like TMs, signal sequences appear to be recognized mostly by lipid partitioning, consistent with their ability to be crosslinked to lipids[26,27], and the correlation between the partitioning of synthetic peptides into hydrophobic solvents and their function as signal sequences *in vivo*[28]. Nevertheless, amino-acid interactions with TM2 of the channel may contribute to the recognition of signal sequences. This may explain why some amino acids occur more frequently than others in signal sequences, even when they have about the same hydrophobicity: leucine is preferred over isoleucine and valine[29], perhaps because its extended side chain can make tighter van der Waals contacts with residues in TM2. Whereas signal sequences of translocating secretory proteins would tend to stay in the binding pocket until they are cleaved off, the more hydrophobic TMs of membrane proteins could move away once the connecting loop to the polypeptide segment inside the channel pore attains adequate length.

While this paper was under review, a cryo-EM structure was published describing an active ribosome-bound mammalian Sec61 channel containing a short secretory polypeptide segment[30]. The authors conclude that the signal sequence helix replaces TM2 of Sec61α in the idle channel, implying that the signal sequence is intercalated into the lateral gate and raising the possibility that the conformational changes differ greatly from those in our system. However, a superposition of the two active channels based on secondary structure matching shows that they are actually very similar, with only moderate differences at the periplasmic/luminal ends of some TMs (Extended Data Fig. 7f). Importantly, in both cases the signal sequence helix docks to the same site outside the lateral gate and runs parallel to TM2 (Extended Data Fig. 7f). Thus, regardless of the organism and mode of translocation, lipid partitioning appears to be the major mechanism by which signal sequences are recognized.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 November 2015; accepted 25 January 2016.
Published online 7 March 2016.

1. Van den Berg, B. *et al.* X-ray structure of a protein-conducting channel. *Nature* **427,** 36–44 (2004).
2. Breyton, C., Haase, W., Rapoport, T. A., Kühlbrandt, W. & Collinson, I. Three-dimensional structure of the bacterial protein-translocation complex SecYEG. *Nature* **418,** 662–665 (2002).
3. Tanaka, Y. *et al.* Crystal structures of SecYEG in lipidic cubic phase elucidate a precise resting and a peptide-bound state. *Cell Reports* **13,** 1561–1568 (2015).
4. Tsukazaki, T. *et al.* Conformational transition of Sec machinery inferred from bacterial SecYE structures. *Nature* **455,** 988–991 (2008).
5. Egea, P. F. & Stroud, R. M. Lateral opening of a translocon upon entry of protein suggests the mechanism of insertion into membranes. *Proc. Natl Acad. Sci. USA* **107,** 17182–17187 (2010).
6. Park, E. *et al.* Structure of the SecY channel during initiation of protein translocation. *Nature* **506,** 102–106 (2014).
7. Gogala, M. *et al.* Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion. *Nature* **506,** 107–110 (2014).
8. Voorhees, R. M., Fernández, I. S., Scheres, S. H. & Hegde, R. S. Structure of the mammalian ribosome-Sec61 complex to 3.4 Å resolution. *Cell* **157,** 1632–1643 (2014).
9. Zimmer, J., Nam, Y. & Rapoport, T. A. Structure of a complex of the ATPase SecA and the protein-translocation channel. *Nature* **455,** 936–943 (2008).
10. Pfeffer, S. *et al.* Structure of the native Sec61 protein-conducting channel. *Nature Commun.* **6,** 8403 (2015).
11. Park, E. & Rapoport, T. A. Mechanisms of Sec61/SecY-mediated protein translocation across membranes. *Annu. Rev. Biophys.* **41,** 21–40 (2012).
12. Park, E. & Rapoport, T. A. Bacterial protein translocation requires only one copy of the SecY complex *in vivo*. *J. Cell Biol*. **198,** 881–893 (2012).
13. Jungnickel, B. & Rapoport, T. A. A posttargeting signal sequence recognition event in the endoplasmic reticulum membrane. *Cell* **82,** 261–270 (1995).
14. Bauer, B. W., Shemesh, T., Chen, Y. & Rapoport, T. A. A "push and slide" mechanism allows sequence-insensitive translocation of secretory proteins by the SecA ATPase. *Cell* **157,** 1416–1429 (2014).
15. Li, W. *et al.* The plug domain of the SecY protein stabilizes the closed state of the translocation channel and maintains a membrane seal. *Mol. Cell* **26,** 511–521 (2007).
16. Harris, C. R. & Silhavy, T. J. Mapping an interface of SecY (PrlA) and SecE (PrlG) by using synthetic phenotypes and *in vivo* cross-linking. *J. Bacteriol.* **181,** 3438–3444 (1999).
17. Tam, P. C., Maillard, A. P., Chan, K. K. & Duong, F. Investigating the SecY plug movement at the SecYEG translocation channel. *EMBO J.* **24,** 3380–3388 (2005).
18. Paetzel, M. Structure and mechanism of *Escherichia coli* type I signal peptidase. *Biochim. Biophys. Acta* **1843,** 1497–1508 (2014).
19. Yamamoto, Y. *et al.* Conformational requirement of signal sequences functioning in yeast: circular dichroism and ^{1}H nuclear magnetic resonance studies of synthetic peptides. *Biochemistry* **29,** 8998–9006 (1990).
20. Rizo, J., Blanco, F. J., Kobe, B., Bruch, M. D. & Gierasch, L. M. Conformational behavior of *Escherichia coli* OmpA signal peptides in membrane mimetic environments. *Biochemistry* **32,** 4881–4894 (1993).
21. Park, E. & Rapoport, T. A. Preserving the membrane barrier for small molecules during bacterial protein translocation. *Nature* **473,** 239–242 (2011).
22. Cannon, K. S., Or, E., Clemons, W. M., Jr, Shibata, Y. & Rapoport, T. A. Disulfide bridge formation between SecY and a translocating polypeptide localizes the translocation pore to the center of SecY. *J. Cell Biol.* **169,** 219–225 (2005).
23. Heinrich, S. U., Mothes, W., Brunner, J. & Rapoport, T. A. The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell* **102,** 233–244 (2000).
24. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–381 (2005).
25. Bischoff, L., Wickles, S., Berninghausen, O., van der Sluis, E. O. & Beckmann, R. Visualization of a polytopic membrane protein during SecY-mediated membrane insertion. *Nature Commun.* **5,** 4103 (2014).
26. Martoglio, B., Hofmann, M. W., Brunner, J. & Dobberstein, B. The protein-conducting channel in the membrane of the endoplasmic reticulum is open laterally toward the lipid bilayer. *Cell* **81,** 207–214 (1995).
27. Plath, K., Mothes, W., Wilkinson, B. M., Stirling, C. J. & Rapoport, T. A. Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell* **94,** 795–807 (1998).
28. McKnight, C. J., Briggs, M. S. & Gierasch, L. M. Functional and nonfunctional LamB signal sequences can be distinguished by their biophysical properties. *J. Biol. Chem.* **264,** 17293–17297 (1989).
29. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10,** 1–6 (1997).
30. Voorhees, R. M. & Hegde, R. S. Structure of the Sec61 channel opened by a signal sequence. *Science* **351,** 88–91 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Martone and for help with nanobody generation and cloning, C. Shoemaker and J. Mukherjee for their assistance in alpaca immunization, A. Whynot for cloning *G. thermodenitrificans* SecY, and H. Suzuki and T. Walz for help with fluorescence size-exclusion chromatography. We thank the staff at Northeastern Collaborative Access Team (NE-CAT) of the Advanced Photon Source, the SBGrid consortium at Harvard Medical School, the organizers of the CCP4/Advanced Photon Source summer school 2015, and the beam host at GMCA-CAT. We thank A. Salic and T. Guettler for reading the manuscript. The work was supported by National Institutes of Health grants to T.A.R. (GM052586) and by a Pioneer Award to H.P. T.A.R. is a Howard Hughes Medical Institute investigator.

Author Contributions E.P. designed the SecA–substrate fusion constructs, performed biochemical tests, and obtained the initial crystals; L.L. optimized constructs and crystals, performed biochemical tests, screened nanobodies, and determined the crystal structures; J.L., J.I., and H.P. generated and cloned nanobodies; T.A.R., L.L., and E.P. interpreted the structure and wrote the manuscript. T.A.R. supervised the project.

Author Information The crystal structure determined in this study has been deposited in the Protein Data Bank (PDB) under accession number 5EUL. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.L. (long_li@hms.harvard.edu) or T.A.R. (tom_rapoport@hms.harvard.edu).

METHODS

Protein translocation and disulfide crosslinking assays. To test *in vivo* OmpA–GFP translocation by *E. coli* SecA and *E. coli* SecY complex (three-component system), *E. coli* strain EP52 or EP62 (ΔrmfΔompT secY-CBP)[12] was transformed with pACYC-SecYEG expressing *E. coli* SecYEG complex with SecY containing a unique Cys at position 68 (ref. 21). The cells also expressed OmpA–GFP from pBAD-OmpA–GFP under an arabinose-inducible promoter[12]. In all constructs, including those used in the two-component system, position −1 of the signal sequence was mutated to Tyr to prevent signal peptide cleavage[31]. The cells were grown at 37 °C to log phase in lysogeny broth (LB) medium supplemented with 100 μg ml^{-1} ampicillin and 40 μg ml^{-1} chloramphenicol. OmpA–GFP expression was induced by addition of 0.15% L-arabinose for 1 h. Where indicated, 0.3 mM copper(II) 1,10-phenanthroline ($CuPh_3$) was added to the bacterial culture for 15 min at room temperature (22 °C). The culture was treated with 10 mM *N*-ethyl maleimide (NEM) for 30 min on ice to block free cysteines. The cells were lysed in SDS sample buffer. Samples (equivalent amounts based on an absorbance at 600 nm, $A_{600\,nm}$, of *E. coli* cultures) were subjected to non-reducing SDS–PAGE and analysed by western blotting with anti-SecY- or GFP-antibodies. Where indicated, the samples were treated with 2% β-mercaptoethanol (β-ME) at 50 °C for 20 min before SDS–PAGE. To test the translocation of *E. coli* SecA-OAIns fusion with *E. coli* SecY complex (two-component system), EP52 cells were transformed with pACYC-SecYEG and pBAD-*Ec*SecA-OAIns. The cells were grown to log phage in LB medium supplemented with 100 μg ml^{-1} ampicillin and 40 μg ml^{-1} chloramphenicol, and SecA-OAIns expression was induced by addition of 0.1% L-arabinose for 4 h at room temperature or 30 min at 37 °C. Disulfide crosslinking and non-reducing SDS–PAGE analysis were performed as described above.

Translocation by *B. subtilis* SecA and *G. thermodenitrificans* SecY complex was tested similarly. In the case of the three-component system, *E. coli* strain EP51 (ΔrmfΔompT)[21] was transformed with pTet-*G. thermodenitrificans* $SecE_{His8}$/Y_{His6}, which expresses bicistronic *G. thermodenitrificans secE* and *secY* genes (SecY contains a Cys at position 60) under a tetracycline-inducible promoter[21]. The cells were also transformed with pBAD-OmpA–GFP/*B. subtilis* SecA, a modified version of pBAD-OmpA, which contains an additional ribosome binding site and the *B. subtilis secA* gene (for bicitronic expression) following OmpA–GFP. In the case of the two-component system, EP51 cells were transformed with pTet-*G. thermodenitrificans* $SecE_{His8}$/Y_{His6} and pBAD-*B. subtilis* SecA-OAIns. The cells were grown to log phase at 37 °C, and the expression of *G. thermodenitrificans* SecYE was induced by addition of 200 ng ml^{-1} anhydrotetracycline for 1.5–2 h at 37 °C. Then, the expression of OmpA–GFP/*Bs*SecA or *Bs*SecA-OAIns was induced by addition of 0.2% L-arabinose for 1–3 h. The cells were collected and subjected to SDS–PAGE followed by western blotting as described above. *G. thermodenitrificans* SecY was detected using anti-His antibodies.

Nanobody library generation. Purified *G. thermodenitrificans* SecYE in *n*-dodecyl β-D-maltoside (DDM)-containing buffer was injected into an alpaca to elicit an immune response. A male alpaca (*Vicugna pacos*) was purchased locally, maintained in pasture, and immunized following a protocol authorized by the Tufts University Cummings Veterinary School Institutional Animal Care and Use Committee. After five rounds of immunization, total RNA was isolated from ~10^6 fresh peripheral blood lymphocytes, using an RNeasy Plus Mini Kit (Qiagen), following the manufacturer's instructions. Total RNA was used to synthesize a complementary DNA (cDNA) library using SuperScript III reverse transcriptase (ThermoFisher Scientific) with a combination of random hexamers, oligo(dT), and gene-specific primers. The variable fragments of heavy chain antibodies (VHHs) segments were further amplified from this cDNA library using primers specific to the VHH region[32]. PCR products were pooled, digested with NotI-HF and AscI (NEB), gel purified, ligated into a M13 phagemid vector (pJSC), and transformed via electroporation into TG1 *E. coli* (Agilent). Library complexity was assessed by serial dilution and plating on 2YT agar plates supplemented with 2% glucose and 10 μg ml^{-1} ampicillin.

Selection of nanobodies by phage display. Purified SecYE and SecA-OAIns/SecYE proteins were biotinylated via coupling to primary amines with a fivefold molar excess of Chromalink NHS biotin reagent (Solulink) for 90 min. Excess biotin reagent was removed using a ZeBa desalting column (ThermoFisher Scientific). Twenty micrograms of each protein were mixed with 100 μl of MyOne Streptavidin T1 Dynabeads (ThermoFisher Scientific) blocked with 2% BSA. The beads were incubated with 200 μl of phage at 10^{13} plaque-forming units per millilitre for 1 h at room temperature. Non-binding phage was washed away and bound phage was eluted first by incubating with 1 ml of saturated ER2738 culture, followed by 200 mM pH 2.2 glycine. The elutions were neutralized, pooled, and plated onto 2YT agar plates supplemented with 2% glucose, 5 μg ml^{-1} tetracycline, and 10 μg ml^{-1} ampicillin. A second round panning was performed with 2 μg of each protein and 40 μl MyOne Streptavidin T1 Dynabeads. All procedures were conducted in 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.02% DDM for SecYE, and with 20 mM HEPES-KOH pH 7.0, 150 mM NaCl, 10% glycerol, 0.02% DDM, 5 mM $MgCl_2$, 0.1 mM ADP/BeFx for SecA-OAIns/SecYE. For each protein, 95 clones were sequenced, and sequences that appeared more than 5 times were selected for subsequent validation.

Nanobody screening. Thirteen distinct families of nanobodies directed against *G. thermodenitrificans* SecYE were identified by DNA sequencing. Twenty-two nanobody clones were sub-cloned into the pHEN6 vector[33], which adds an N-terminal pelB sequence and C-terminal sortase and His_6 tags (LPETGG-His_6). The proteins were expressed in 5 ml *E. coli* cultures. After Ni-resin purification, the nanobodies were labelled with Alexa 555 (Invitrogen) by sortase reaction[34]. The labelled nanobodies (1 μM concentration) were mixed with *G. thermodenitrificans* SecA-OAIns/SecYE complex or with SecYE alone at a molar ratio of 2:1. Nanobody binding was monitored by a shift of the peak in size-exclusion chromatography coupled with a fluorescence detector (Shimadzu). All 22 nanobodies bound to SecYE, but only nanobody AYC08 had a high affinity for SecA-OAIns/SecYE. The binding of AYC08 to free SecYE was weaker than to SecA-OAIns/SecYE, indicating that AYC08 interacts with the periplasmic side of SecY and that the binding epitope is only fully exposed in the active channel.

Protein expression and purification. *E. coli* strain EP51 was transformed with pTet-*G. thermodenitrificans* $SecE_{His8}$/Y and pBAD-*B. subtilis* SecA-OAIns49(L7). Residues 202–213 in the loop between TM5 and TM6 of SecY were replaced by the sequence TFGGLN. Cells were grown in LB medium supplemented with 100 μg ml^{-1} ampicillin, 40 μg ml^{-1} chloramphenicol, and 0.5% glycerol at 37 °C until $A_{600\,nm}$ reached 0.6–0.7. The expression of the *G. thermodenitrificans* SecYE was induced by addition of 200 ng ml^{-1} anhydrotetracycline, and cells were incubated for 1.5 h at 37 °C and an additional 1 h at 22 °C. Then 0.15% L-arabinose was added to the culture to express *B. subtilis* SecA-OAIns overnight at 16 °C. Cells were harvested by centrifugation and stored at −80 °C until use.

The cells were suspended in buffer A (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM phenylmethanesulfonyl fluoride) and lysed in a microfluidizer (Microfluidics). The membranes were pelleted by ultracentrifugation, washed once with buffer B (20 mM Tris-HCl pH 7.5, 300 mM NaCl), and solubilized with 1% DDM (Anatrace) in buffer C (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol). After 1 h incubation at 4 °C, the solution was clarified by ultracentrifugation. The supernatant was loaded onto a 5 ml POROS-MC20 column (Applied Biosystems) pre-charged with $CoCl_2$. After washing with 15 ml of buffer D (as buffer C, but with 0.02% DDM) containing 10 mM imidazole and 5 ml of buffer D containing 15 mM imidazole, the protein was eluted with 5.5 ml of buffer D containing 250 mM imidazole. Immediately after elution, 1 mM ADP/BeFx, 5 mM $MgCl_2$, and 0.1 mg ml^{-1} *E .coli* polar lipids (25 mg ml^{-1} stock dissolved in 1% DDM) were added. To cleave the GFP-strep tag, 3C protease was added at a ratio of 1:30 (w:w) and the mixture was incubated overnight at 4 °C. The sample was diluted 1:1 (v:v) with buffer E (20 mM Tris-HCl pH 7.5, 10% glycerol, 0.02% DDM, 5 mM $MgCl_2$) and loaded onto a Mono Q 10/100 column (GE Healthcare). The protein was eluted with a gradient of 15–35% buffer F (20 mM Tris-HCl pH 7.5, 1 M NaCl, 10% glycerol, 0.02% DDM, 5 mM $MgCl_2$). The peak fractions were collected, and 0.1 mg ml^{-1} *E .coli* polar lipids and 1 mM ADP/BeFx were added. The protein was concentrated with an Amicon filter (100 kDa MWCO, EMD Millipore) and loaded onto a Superdex 200 10/300 column (GE Healthcare) in buffer G (20 mM HEPES-KOH pH 7.0, 100 mM NaCl, 10% glycerol, 0.02% DDM, 5 mM $MgCl_2$, 1 mM ADP/BeFx). The peak fractions were concentrated to ~12 mg ml^{-1}, aliquoted, and flash-frozen in liquid nitrogen. The protein was stored at −80 °C and thawed right before crystallization.

The plasmid coding for nanobody AYC08 was transformed into WK6 cells. The cells were grown in 2xYT medium at 37 °C and protein expression was induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at $A_{600\,nm} = 0.6$. The incubation was continued overnight at 30 °C. About 5 g of cells were obtained from 1 L of culture. After suspension in 15 ml of TES buffer (200 mM Tris-HCl pH 8.0, 500 μM EDTA, 500 mM sucrose), the sample was diluted 1:3 (v:v) in ice-cold water for 3 h to induce cell lysis. After centrifugation, the supernatant was mixed with 5 ml of Ni-NTA resin (Qiagen) and incubated at 4 °C for 1 h. The protein was eluted with 250 mM imidazole. The protein was further purified by gel filtration on a Superdex 200 10/300 column in 20 mM HEPES-KOH pH 7.0, 100 mM NaCl, 10% glycerol, 5 mM $MgCl_2$). The purified nanobody was concentrated to 30 mg ml^{-1}.

Crystallization. Initial crystals were obtained with a complex of *G. thermodenitrificans* SecYE and either *B. subtilis* SecA-OAIns and or *B. subtilis* SecA-DsbAIns (the last containing the signal sequence of DsbA instead of that of OmpA). Only small crystals were obtained and diffracted to a maximum resolution of 6.5 Å at the synchrotron X-ray source. A variety of approaches were tested to improve the crystals, such as inserting different hydrophilic proteins into each of the periplasmic loops of SecY (T4 lysozyme[35], cytochrome *b*-562 (ref. 36), P1/P4 domain of SecD/F[37], ROP helical bundle[38]), truncating SecY loops, using various detergents, co-expressing SecG, and employing Fab-fragments

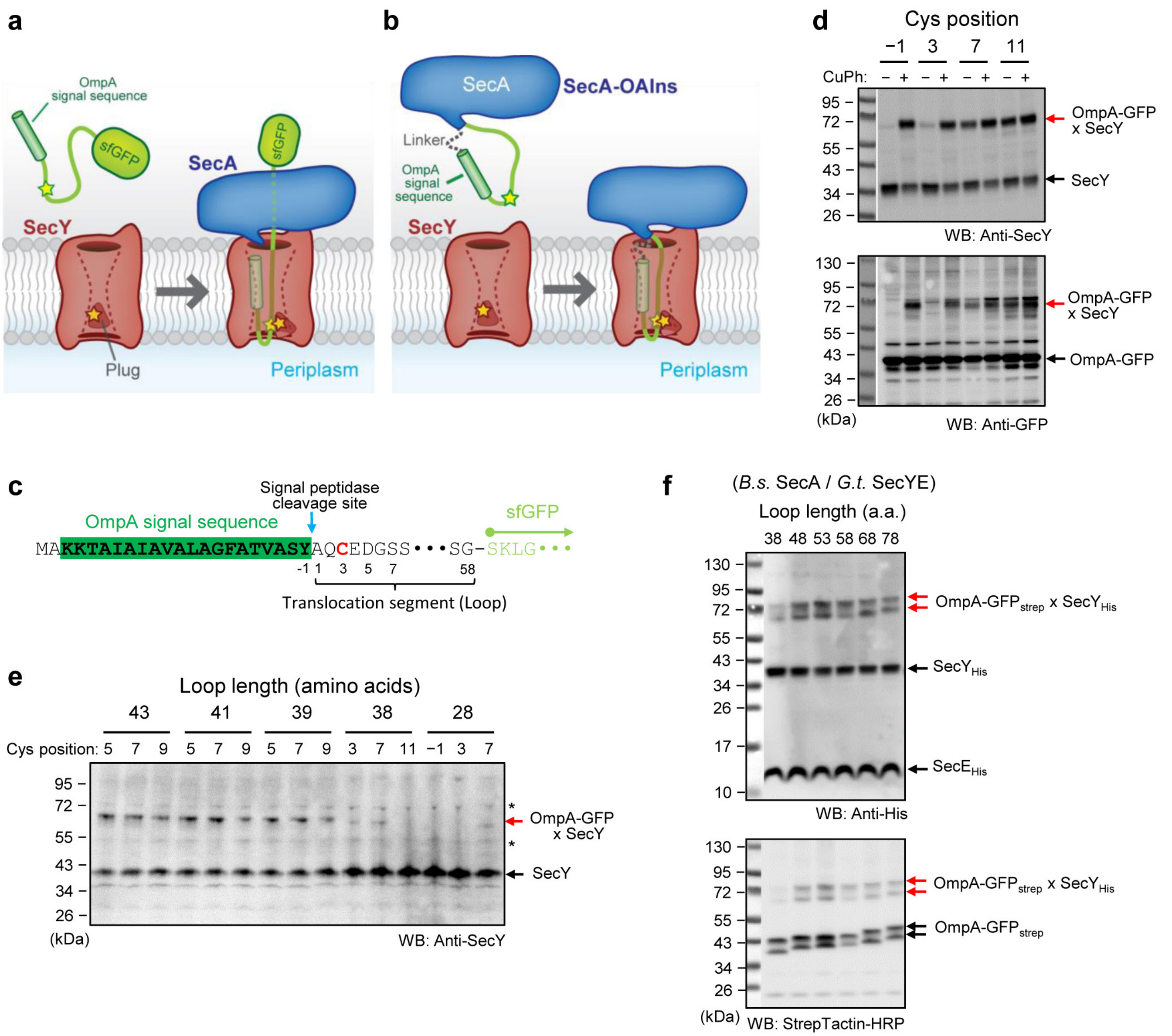
of monoclonal antibodies generated against SecY. However, crystals with improved diffraction were only obtained when the complex was co-crystallized with nanobody AYC08.

The complex of *G. thermodenitrificans* SecYE and *B. subtilis* SecA-OAIns was mixed with nanobody AYC08 at a molar ratio of 1:1.2 with addition of 1 mg ml^{-1} lipids (42 mg ml^{-1} 1,2-dipalmitoyl-*sn*-glycero-3-phosphoglycerol (DPPG) plus 1,2-dipalmitoyl-*sn*-glycero-3-phosphoethanolamine (DPPE) (3:1) suspension in 0.5% DDM)[39]. The mixture was incubated at 4 °C overnight and clarified by ultracentrifugation before setting up crystallization trays. The initial crystal screening yielded several crystal forms. Three of them were readily reproducible, but all diffracted to a maximum of ~6 Å resolution. Heavy atom compounds were screened for crystal soaking and the Ta_6Br_{12} cluster improved the resolution limit of one crystal form. The best crystals were obtained with the hanging drop method, mixing 0.5 μl of the protein solution and 0.5 μl of well solution (21–24% polyethylene glycol (PEG) 1500, 100 mM Tris-HCl pH 8.5, 50–100 mM $MgAc_2$, 2% 2-methyl-2,4-pentandiol; MPD) and using 24-well VDX plates with 500 μl well solution. The crystals were grown at 22 °C over a week. The Ta_6Br_{12} powder (Jena Bioscience) was suspended in buffer at a concentration of 20 mM and added to the crystallization drops at a final concentration of ~2 mM. After overnight incubation, the crystals were flash-frozen in liquid nitrogen.

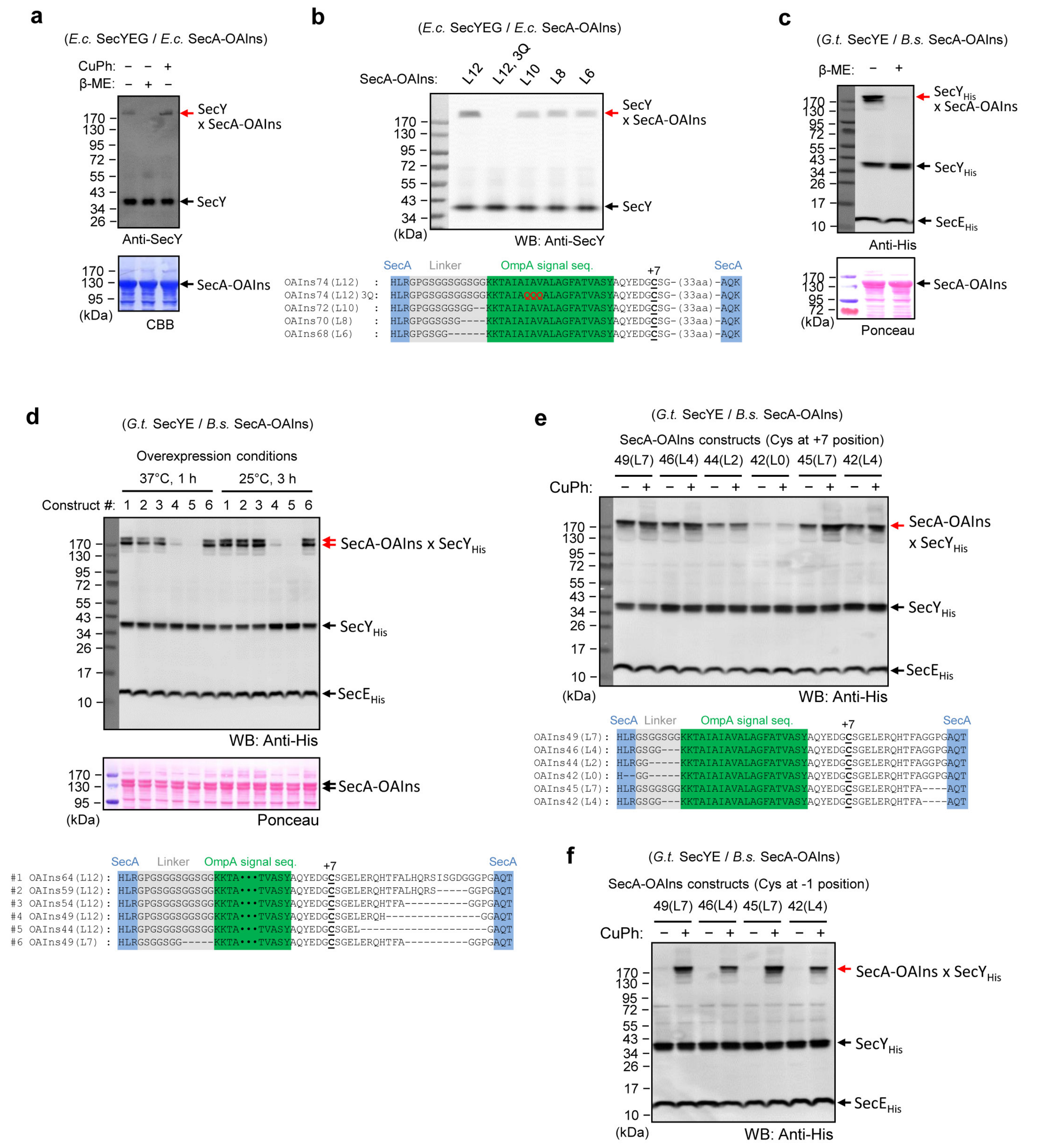
X-ray data collection and structure determination. Hundreds of crystals were screened at NE-CAT and GM/CA-CAT of the Advanced Photon Source (Argonne National Laboratory). The diffraction of the crystals decayed rapidly and was weak, caused by strong heavy atom absorption. However, a complete set of three-wavelength MAD data from a single crystal could be collected at GM/CA-CAT. The data were processed with the XDS package[40] and analysed by the program Aimless[41]. The crystal belongs to the $P6_122$ space group. The diffraction was anisotropic. Along axis c^*, the diffraction went beyond 3.70 Å (I/sigma = 2.0), whereas along the axes a^* and b^*, the diffraction was limited to 4.48 Å (I/sigma = 1.5). The data were initially processed to 3.9 Å. Anisotropy correction was applied in the different programs used in the following calculations. Molecular replacement was used to locate SecA, SecYE, and the nanobody, employing as search models *B. subtilis* SecA (PDB accession number 1TF5), *T. maritima* SecYE (PDB accession number 3DIN), and an anti-DHFR nanobody (PDB accession number 4EIG), respectively. The crystal contained one complex per asymmetric unit with a solvent content of 69%. Heavy atom sites were identified with the molecular replacement results by MR-SAD in Phaser[42] and refined without molecular replacement models in Sharp (Global Phasing). The experimental map based on the positions of the heavy metal ion clusters contained useful phase information to 5.5 Å (judged by phasing power of 1). The overall figure of merit was 0.44 (acentric) and 0.42 (centric). The phases were extended to 3.9 Å through density modification, using the programs Resolve[43] and CNS[44]. After density modification, most of the α-helices were well resolved and some of the large side chains were visible in B-factor-sharpened maps. To improve the density map, model phases were combined with experimental phases. Models for SecA and the nanobody molecules were placed first and modified according to the density map. The SecY and SecE molecules were then built with the guidance of B-factor-sharpened maps and the crystal structures of *T. thermophilus*[4] and *T. maritima* SecYE[9]. A density map for tracing the translocating peptide was generated by combining phase information from MAD phasing, and models of SecA, nanobody, and SecYE (Extended Data Fig. 4c). The signal sequence was initially modelled as an ideal poly-Ala helix and placed into the density map. The registry of the signal sequence helix was then determined from the density for two aromatic residues (Phe(−7) and Tyr(−1)). The following 23 amino acids of the translocating peptide were traced in a B-factor-sharpened density map. The registry of this segment was determined on the basis of density for an aromatic residue (Phe3), several negatively charged residues (Glu4, Asp5, Glu10, and Glu12) surrounding the positively charged Ta_6Br_{12} cluster, and the cysteine engaged in the disulfide bond (Cys7). Our model places Gly19 inside the pore ring, but it is possible that the registry is off by one residue (Ala18 or Gly20 would be in the pore), an ambiguity that does not alter the interpretation of the model. Annealing to a temperature of 2,500 K was applied to the model at an early stage of refinement with the program Phenix[45]. The individual XYZ and group B factors were refined by using both the Phenix and Refmac5 (ref. 46) programs. Secondary structure was tightly constrained during refinement. The single-wavelength anomalous diffraction (SAD) likelihood function implemented in Refmac5 was used to refine the heavy atom clusters together with the protein model. The model was improved by molecular replacement and multiple cycles of model building and refinement. Manual adjustments were made in COOT[47] and refinement was performed in reciprocal space. At a late stage of refinement, the diffraction data were extended to 3.7 Å, which helped to resolve some regions in the density map, for example of the translocating peptide. We also tested the data with ellipsoidal truncation processed by the Anisotropy server (http://services.mbi.ucla.edu/anisoscale/). The truncated data produced better density maps, which were used for refinement as well. The final model was refined to an R_{work} value of 30.6% and an R_{free} value of 32.5% and showed good geometry (Extended Data Table 1). An attempt was made to include the model-refined Ta_6Br_{12} clusters as the resolved heavy atom substructure for recalculation of SAD and MAD experimental phases. However, this did not improve the map, in part because individual metal atoms of the clusters could not be accurately positioned, as observed in other cases of similar resolution[48]. While the centres of the clusters are well defined, the individual metal atom positions in our model should be considered to be very approximate. To ensure that the R-factor is not dominated by the heavy metal ion clusters, we tested its sensitivity to changes in the protein model. R_{free} increased by 0.013–0.017 upon deletion of any of the TMs of SecY or of the signal sequence helix. A similar increase was observed when the same analysis was performed for three different membrane proteins of similar size (2ZD9, 4CZB, 4CDI; 1,078–1,655 amino acids), the structure of which was determined in the absence of heavy metal ion clusters at similar resolution (3.5–4 Å), solvent content (0.63–0.85), and R-factor (0.3–0.34). Model validation was performed by using PHENIX. The following regions could not be traced: residues 1–15, 244–247, 262–264, 271–272, 620–626, and 635–712 of SecA; residues 1–12, 145–146, 200–213, 244–260, and 291–308, 390–391, and 396–398 of SecY; residues 1–3, 22–23, 58–60 of SecE; residues 30–31, 42, and 100 of the nanobody; and the first five residues of the linker preceding the signal sequence. In addition, some amino-acid side chains were not well resolved, so they were modelled as Ala. Some density close to metal ion cluster sites 15–18 remains unexplained. Figures showing the structures were generated with Chimera[49]. All the X-ray crystallographic software was maintained by SBGrid[50].

Accession numbers of structures used in this paper. The accession numbers used are PDB 1RH5, 3MP7, 2ZJS, 3DIN, 5AWW, 3JC2, 4CG6. The new structure described in this paper has accession number 5EUL.

31. Fikes, J. D., Barkocy-Gallagher, G. A., Klapper, D. G. & Bassford, P. J. Jr. Maturation of *Escherichia coli* maltose-binding protein by signal peptidase I *in vivo*. Sequence requirements for efficient processing and demonstration of an alternate cleavage site. *J. Biol. Chem.* **265,** 3417–3423 (1990).
32. Maass, D. R., Sepulveda, J., Pernthaner, A. & Shoemaker, C. B. Alpaca (*Lama pacos*) as a convenient source of recombinant camelid heavy chain antibodies (VHHs). *J. Immunol. Methods* **324,** 13–25 (2007).
33. Arbabi Ghahroudi, M., Desmyter, A., Wyns, L., Hamers, R. & Muyldermans, S. Selection and identification of single domain antibody fragments from camel heavy-chain antibodies. *FEBS Lett.* **414,** 521–526 (1997).
34. Guimaraes, C. P. *et al.* Site-specific C-terminal and internal loop labeling of proteins using sortase-mediated reactions. *Nature Protocols* **8,** 1787–1799 (2013).
35. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).
36. Chun, E. *et al.* Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* **20,** 967–976 (2012).
37. Tsukazaki, T. *et al.* Structure and function of a membrane component SecDF that enhances protein export. *Nature* **474,** 235–238 (2011).
38. Hari, S. B., Byeon, C., Lavinder, J. J. & Magliery, T. J. Cysteine-free Rop: a four-helix bundle core mutant has wild-type stability and structure but dramatically different unfolding kinetics. *Protein Sci.* **19,** 670–679 (2010).
39. Gourdon, P. *et al.* HiLiDe—systematic approach to membrane protein crystallization in lipid and detergent. *Cryst. Growth Des.* **11,** 2098–2106 (2011).
40. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
41. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D* **69,** 1204–1214 (2013).
42. McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. D* **63,** 32–41 (2007).
43. Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr. D* **56,** 965–972 (2000).
44. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2,** 2728–2733 (2007).
45. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
46. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67,** 355–367 (2011).
47. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
48. Banumathi, S., Dauter, M. & Dauter, Z. Phasing at high resolution using Ta_6Br_{12} cluster. *Acta Crystallogr. D* **59,** 492–498 (2003).
49. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
50. Morin, A. *et al.* Collaboration gets the most out of software. *eLife* **2,** e01456 (2013).
51. Pédelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnol.* **24,** 79–88 (2006).

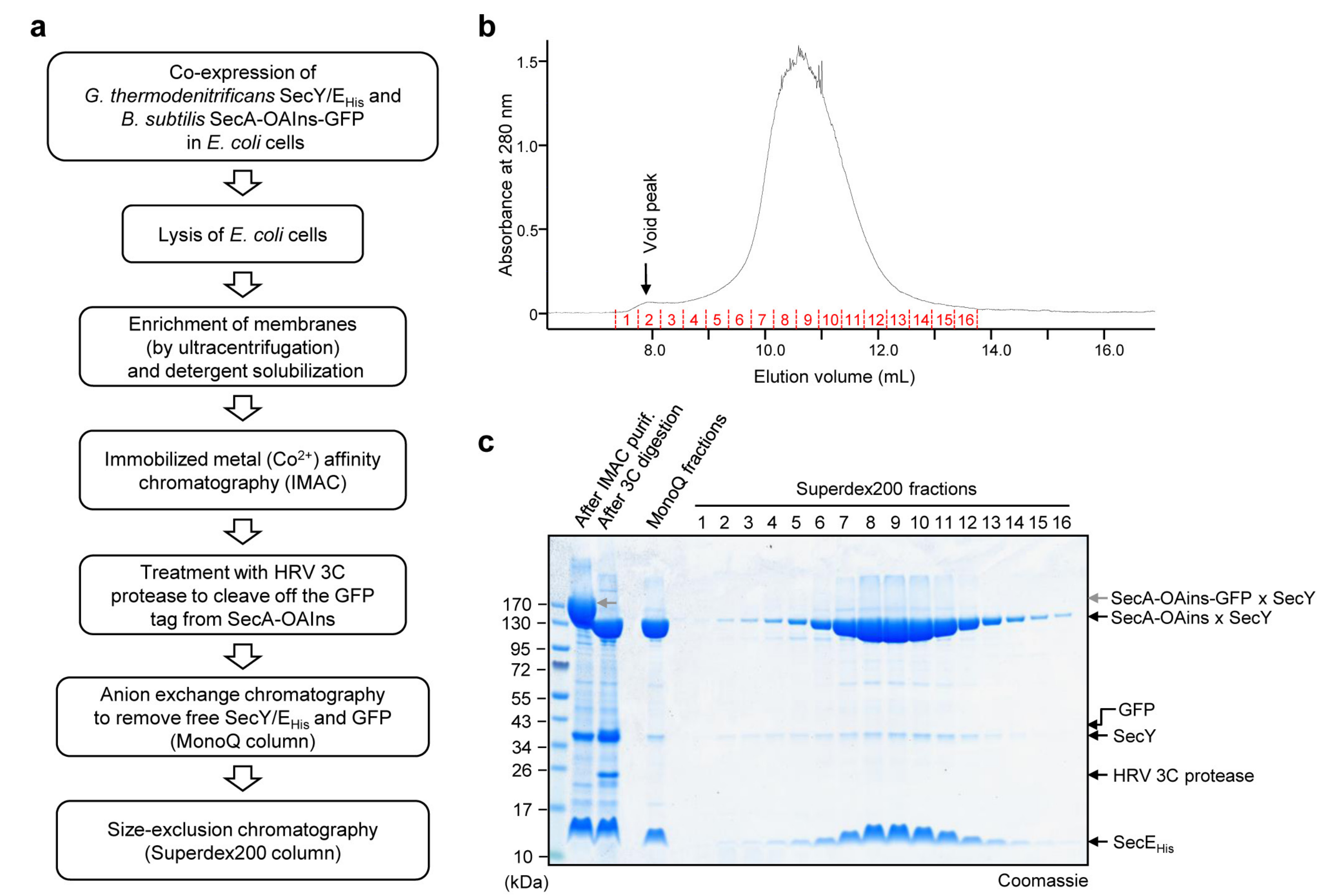


Extended Data Figure 1 | Generation of translocation intermediates with a three-component system. a, Strategy to generate SecA-dependent translocation intermediates in *E. coli* cells. The intermediates are assembled from *E. coli* SecA, *E. coli* SecY complex, and substrate containing an N-terminal OmpA signal sequence and C-terminal superfolder GFP[51] (sfGFP). After loop insertion into the SecY channel, translocation of the C terminus is stalled by the folded sfGFP. Insertion is monitored by disulfide bond formation between a pair of cysteines introduced into the substrate and the plug of SecY (yellow stars). **b**, Scheme showing the simplified system, in which a secretory protein segment is fused into the two-helix finger of SecA. **c**, Sequence of the substrate used in **a**. The −1 position of the original signal sequence was changed to Tyr to prevent signal sequence cleavage. The position of the cysteine and the length of the translocated segment were varied (here shown for Cys at position +3 and 58 amino acids in length). **d**, Variation of the Cys position with a translocationg segment of 58 residues. Where indicated, disulfide crosslinks to SecY with a Cys at position 68 (OmpA–GFPxSecY) were induced by the oxidant copper phenanthroline (CuPh) before harvesting the cells. The samples were analysed by non-reducing SDS–PAGE, followed by western blotting (WB) with anti-SecY and anti-GFP antibodies. **e**, As in **d**, but in the absence of oxidant, with Cys at different positions and variation of the length of the translocated segment. Asterisks indicate non-specific bands. **f**, As in **e**, but with *E. coli* cells expressing *B. subtilis* (*B.s.*) SecA and *G. thermodenitrificans* (*G.t.*) SecYE. The substrate contained a Cys at position +7, and SecY a Cys at position 60. The red arrows indicate spontaneously generated disulfide crosslinks (GFP sometimes does not unfold in SDS, resulting in two bands). The OmpA–GFP constructs contained a C-terminal Strep-tag that was detected by StrepTactin conjugated with horseradish peroxidase (HRP).

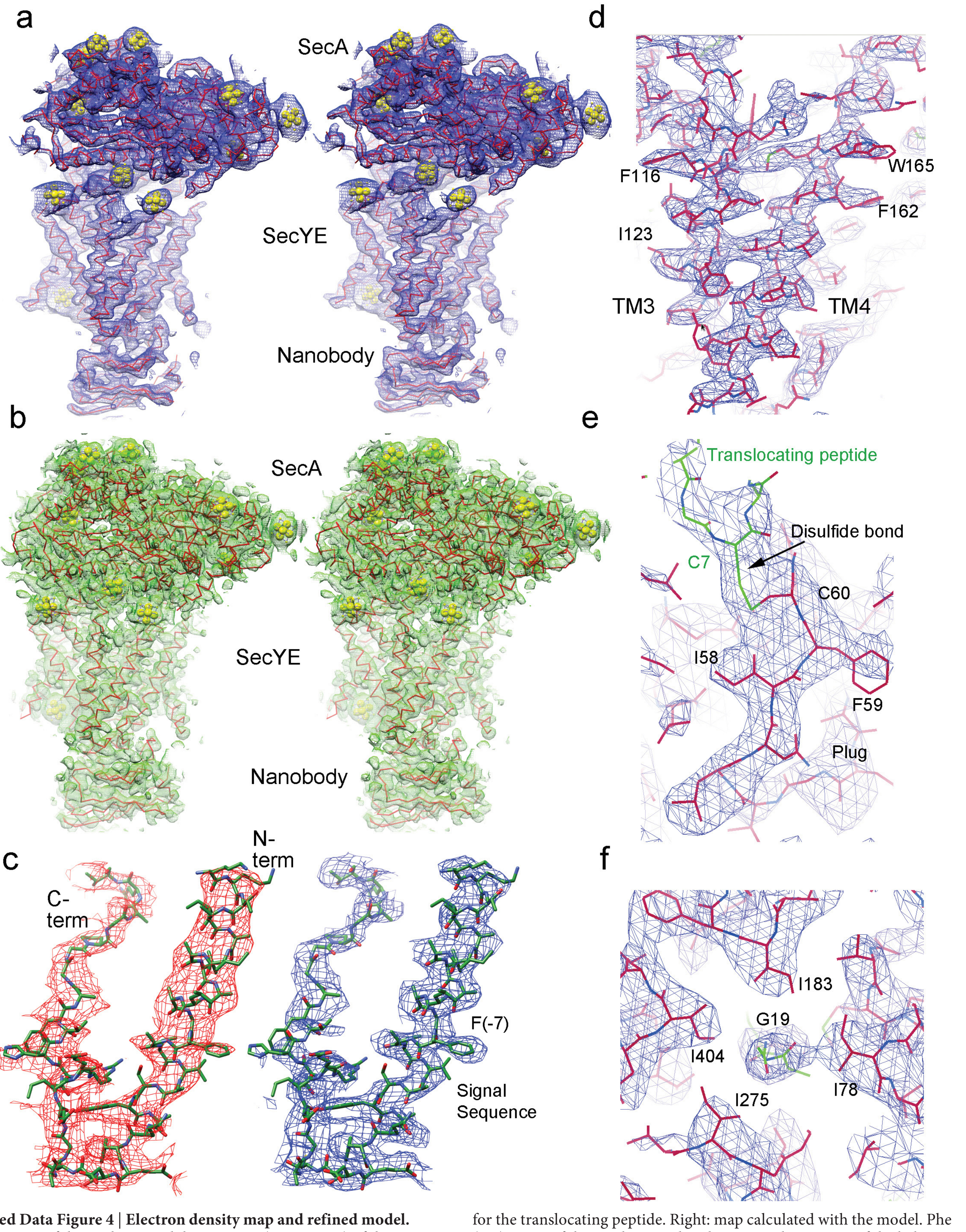


Extended Data Figure 2 | See next page for caption.

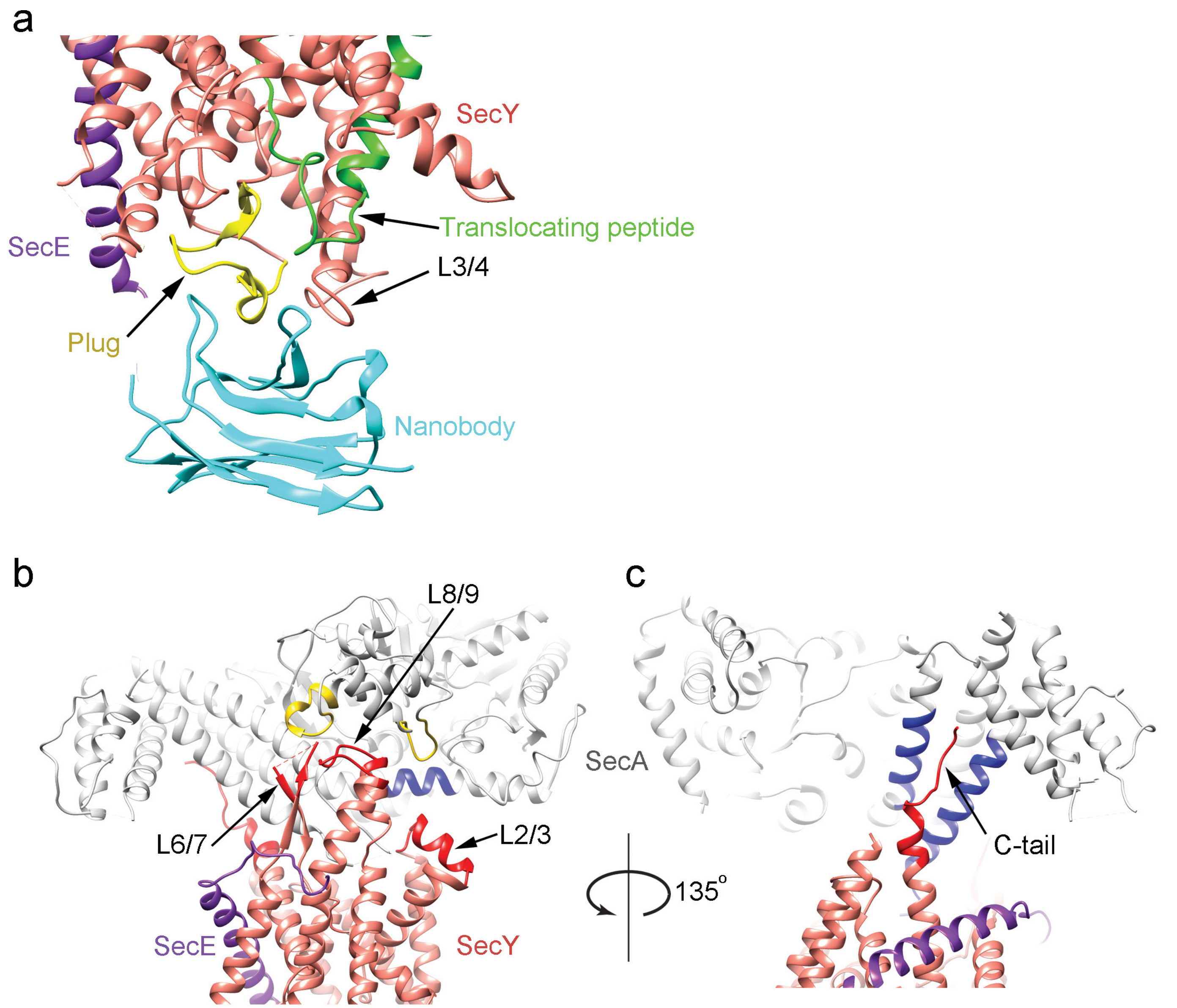
Extended Data Figure 2 | Generation of translocation complexes with SecA–substrate fusion constructs. a, Translocation complexes were generated as indicated in the scheme in Extended Data Fig. 1b. An *E. coli* SecA–substrate fusion (SecA-OAIns74 (L12)) was overexpressed together with *E. coli* SecY complex in *E. coli* cells. SecA-OAIns74 (L12) contains 74 amino acids inserted into the two-helix finger of SecA, including a linker of 12 residues, and a GFP tag following SecA. Translocation of the substrate segment was monitored by spontaneous disulfide crosslinking between a cysteine at position +7 (with respect to the original signal peptidase cleavage site) and a cysteine at position 68 in the plug of SecY. Where indicated, β-mercaptoethanol (β-ME) was added to reduce the disulfide bond. The samples were analysed by non-reducing SDS–PAGE and western blotting with anti-SecY antibodies. The overexpression of SecA-OAIns was monitored by the fluorescence of GFP (data not shown) and staining with Coomassie blue (CBB, lower panel). **b**, As in **a**, but with *E. coli* SecA-OAIns constructs containing from 6 to 12 residues in the linker (L6–L12) or mutations (3Q) in the H-region of the signal sequence. Expression of SecA-OAIns was verified by the strong green fluorescence of cell lysates, caused by GFP fused to the C terminus of SecA (not shown). The lower panel shows the sequences of the SecA-inserted segments. **c**, As in **a**, but with *B. subtilis* SecA-OAIns74 (L12) and *G. thermodenitrificans* SecYE. SecY and SecA were detected by western blotting with anti-His antibodies and Ponceau staining, respectively. **d**, As in **c**, but with *B. subtilis* SecA-OAIns containing different inserted segments. SecA-OAIns was expressed under different conditions, as indicated. Expression of SecA-OAIns was verified by green fluorescence of GFP fused to the C terminus of SecA (not shown) and Ponceau staining (second panel). The sequences of the constructs are shown in the lowest panel. **e**, As in **d**, but with different constructs, the sequences of which are shown in the lower panel. Where indicated, copper phenanthroline (CuPh) was added to the cells to induce disulfide bridge formation. SecA-OAIns49(L7) was used for crystallization. **f**, As in **e**, but with a Cys at position −1 (the last residue of the OmpA signal sequence) instead of position +7. Note that in this case disulfide formation does not occur spontaneously.



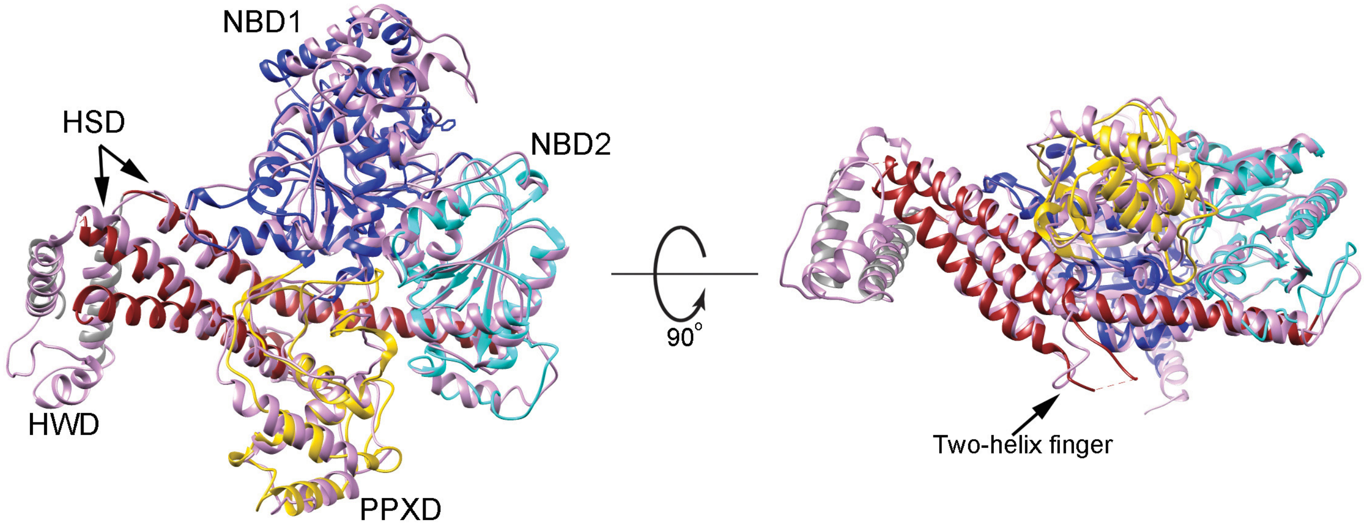
Extended Data Figure 3 | Purification of an active translocation complex. a, Scheme of the purification protocol. **b**, Elution of the *G. thermodenitrificans* SecYE/*B. subtilis* SecA-OAIns complex from a Superdex200 column during the last chromatography step. **c**, Non-reducing SDS–PAGE analysis of samples taken during the purification procedure and of fractions indicated with red numbers in **b**. Lane 1, molecular mass markers. Lane 2, sample analysed after immobilized metal ion affinity chromatography. Lane 3, sample after cleavage of the GFP tag. Lane 4, sample after anion exchange chromatography (MonoQ).



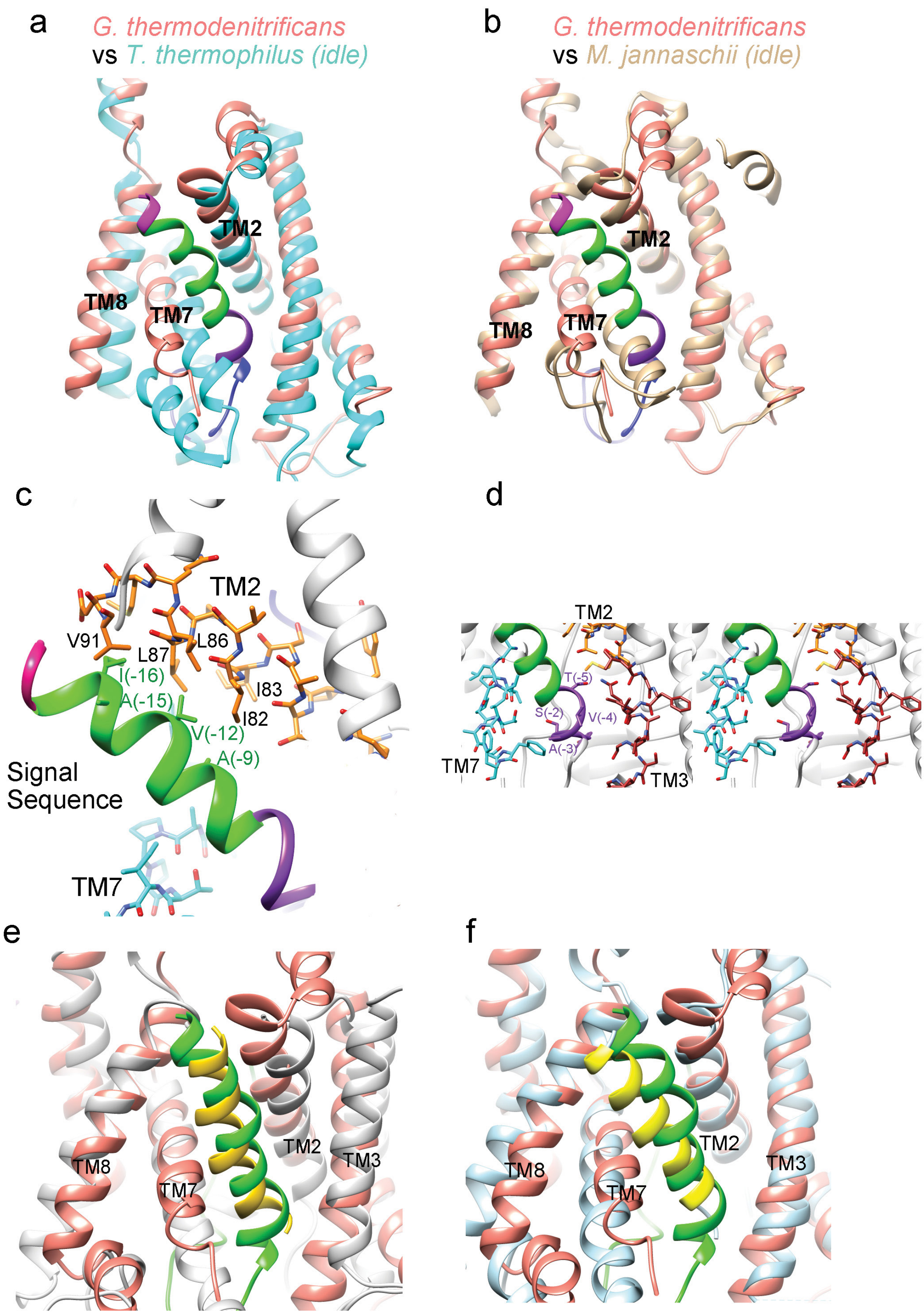
Extended Data Figure 4 | Electron density map and refined model. **a**, Stereo view of the unsharpened density map ($2F_o - F_c$; 1σ) of the entire complex. Heavy metal ion clusters are shown in yellow. **b**, As in **a**, but with the density map derived from MAD phasing after density modification. **c**, SigmaA-weighted phase-combined $2F_o - F_c$ density maps of the translocating peptide region. Left: omit map calculated without a model for the translocating peptide. Right: map calculated with the model. Phe (−7) is one of the residues used to determine the registry of the helix. **d**, A side view of the density for TM3 and TM4. **e**, Density showing the disulfide crosslink between the plug and translocating chain. **f**, Top view of Gly19 of the translocating chain surrounded by pore residues.



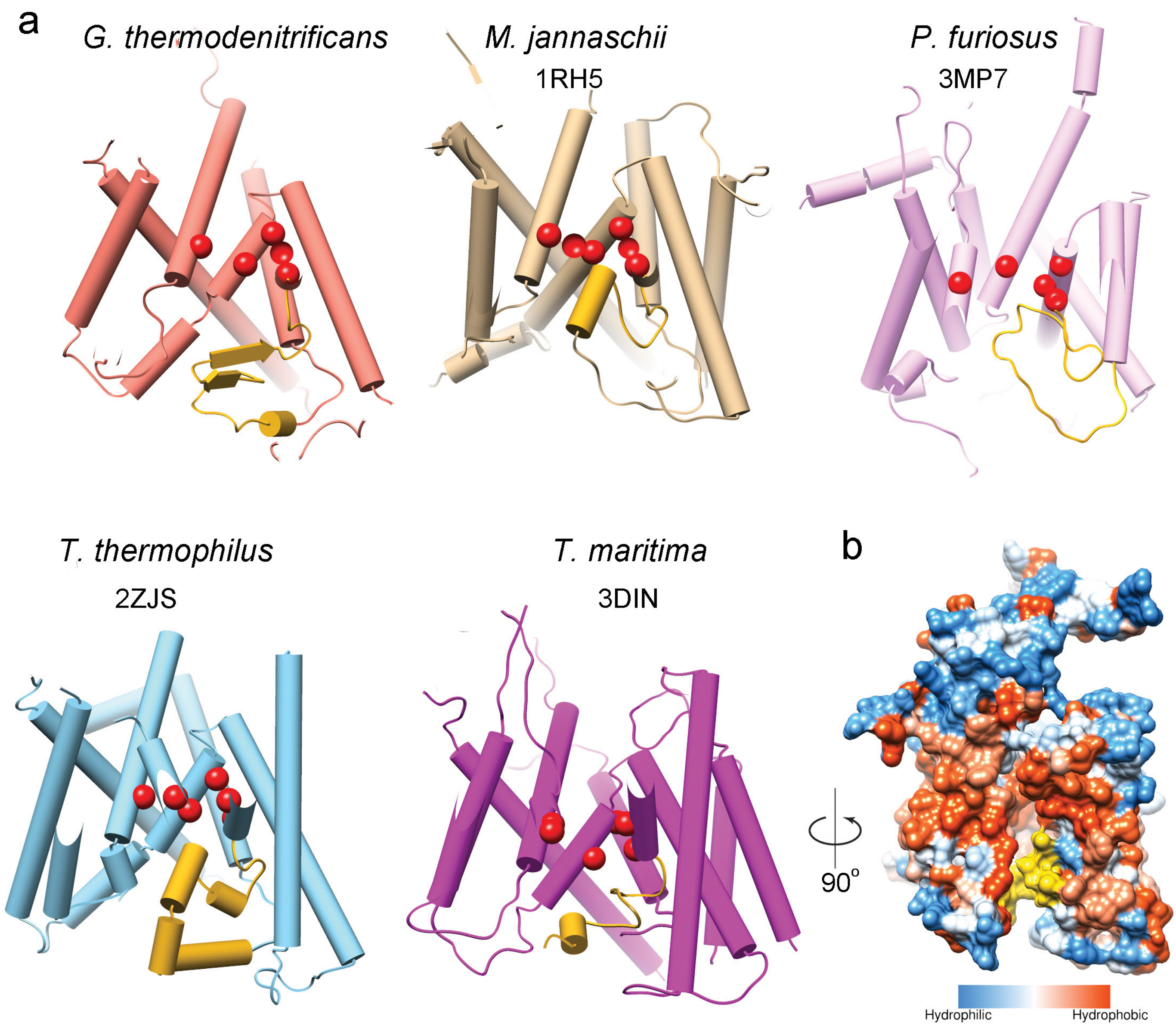
Extended Data Figure 5 | Interactions of the nanobody and SecA with SecY. a, The nanobody binds to the plug and to the loop between TM3 and TM4 (L3/4). **b**, The polypeptide crosslinking domain (PPXD; in yellow) of SecA interacts with the loop between TM8 and TM9 of SecY (L8/9; in red), and the long helix of the helical scaffold domain (HSD; in blue) with the loop between TM2 and TM3 (L2/3; in red). The loop between TM6 and TM7 of SecY (L6/7; in red) does not seem to make strong contact with SecA. **c**, Two helices of the HSD interact with the C-terminal tail of SecY (C-tail; in red).



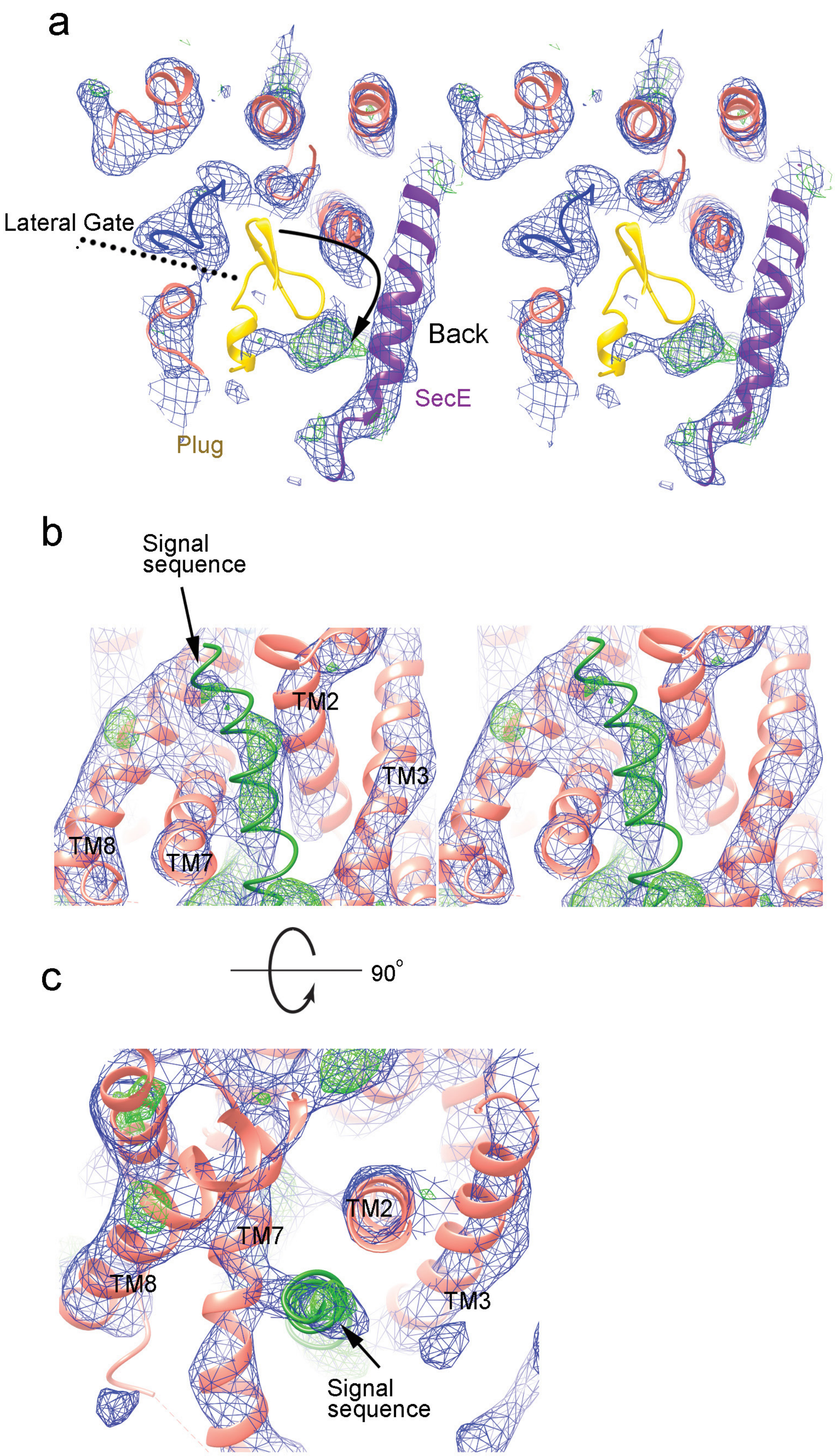
Extended Data Figure 6 | Comparison of the conformations of SecA in the active *G. thermodenitrificans* and inactive *T. maritima* complexes. The domains of SecA in the *G. thermodenitrificans* complex are labelled with different colours (nucleotide binding domain 1 (NBD1), blue; nucleotide binding domain 2 (NBD2), cyan; helical scaffold domain (HSD), brown; helical wing domain (HWD), grey; polypeptide crosslinking domain (PPXD), yellow). SecA in the *T. maritima* complex is shown in pink. Left: a top view (the channel would be underneath); right: a side view with the two-helix finger (part of helical scaffold domain) indicated.



Extended Data Figure 7 | Localization of signal sequences in the *G. thermodenitrificans* SecY and mammalian Sec61 channels and of a TM domain in the mammalian Sec61 channel. a, In the active channel (salmon), the signal sequence displaces TM7 and TM8 in the idle *T. thermophilus* channel (cyan). **b**, As in **a**, but comparison with the idle *M. jannaschii* channel (tan). The C-region of the signal sequence takes the position of TM7. **c**, Side view of the interactions of the H-region of the signal sequence with TM2 of *G. thermodenitrificans* SecY. Interacting amino acids are indicated. **d**, Stereo view showing the intercalation of the C-region into the periplasmic side of the lateral gate. Residues of the amphipathic helix are indicated. **e**, The active *G. thermodenitrificans* channel (in salmon) was aligned with a mammalian channel (in grey) containing a nascent membrane protein (PDB accession number 4CG6) using secondary structure matching[47]. The signal sequence in the bacterial channel is shown in green, and the TM segment of the nascent membrane protein in yellow. **f**, As in **e**, but comparison of the active *G. thermodenitrificans* channel with a mammalian Sec61 channel (light blue) containing a secretory protein fragment (PDB accession number 3JC2). The signal sequences are shown in green and yellow, respectively.



Extended Data Figure 8 | The plug domains in SecY channels. a, The plugs in channels of different organisms have different structures. Shown are side views with the plugs in yellow and pore residues as red spheres. PDB accession numbers are given below the names of the organisms. For the *G. thermodenitrificans* channel, the translocating peptide segment was omitted. **b**, In the inactive *T. maritima* SecY channel, the plug (in yellow) is at the front of the channel, partly sealing the periplasmic side of the lateral gate. Shown is a side view in a surface representation, with hydrophilic and hydrophobic residues in blue and orange, respectively.



Extended Data Figure 9 | Structures of the active *G. thermodenitrificans* complex determined without nanobody or with a different signal sequence. **a**, Stereo view of density maps at 6.5 Å resolution for the active complex in the absence of nanobody. Shown is a $2F_o - F_c$ density map at 1σ (blue mesh) and a difference map ($F_o - F_c$) at 3σ (green mesh), both calculated by molecular replacement with a model lacking the plug. Strong positive density is seen close to SecE, probably corresponding to parts of the plug. The arrow indicates the movement of the plug from the position in the structure with nanobody to the density seen in the structure without nanobody. **b**, Stereo views of density maps at ~8.5 Å resolution for the active *G. thermodenitrificans* complex in which the OmpA signal sequence was replaced by that of DsbA. Shown is a side view of the $2F_o - F_c$ density map at 1σ (blue mesh) and a difference map ($F_o - F_c$) at 3σ (green mesh), both calculated by molecular replacement with a model lacking the signal sequence. Note that the model for the OmpA signal sequence fits well into the density corresponding to the DsbA signal sequence. **c**, As in **b**, but top view and not in stereo.

Extended Data Table 1 | Data collection and refinement statistics for MAD structures

		SecA-OAIns/SecYE + Ta_6Br_{12}	
Data collection		GM/CAT-CAT @APS	
Space group		$P6_122$	
Cell dimensions			
a, *b*, *c* (Å)		127.798,127.798,554.772	
α, β, γ (°)		90, 90, 120	
	Remote	*Inflection*	*Peak*
Wavelength	1.2782	1.2552	1.2548
Resolution (Å)	3.70 (3.80-3.70)	4.00 (4.10-4.00)	4.50 (4.62-4.50)
R_{sym} or R_{merge}	0.084 (>1)	0.095 (>1)	0.128 (>1)
$I/\sigma I$	9.30 (0.4)	9.69 (0.9)	11.90 (1.0)
Completeness (%)	99 (100)	99 (100)	99 (100)
CC1/2	0.999 (0.57)	0.998 (0.80)	0.999 (0.92)
Redundancy	10.6 (10.2)	10.6 (10.4)	21.3 (21.6)
Refinement			
Resolution (Å)	3.7		
No. reflections	53812		
$R_{work/}$ R_{free}	0.306/0.325		
No. atoms	9888		
Protein	9532		
Ligand/ion	356		
Water	N/A		
B-factors	194.00		
Protein	190.24		
Ligand/ion	300.34		
Water	N/A		
R.m.s deviations			
Bond lengths (Å)	0.007		
Bond angles (°)	1.15		
Ramachandran			
Favored / allowed / outliers (%)	91.3 / 6.8 / 1.9		

*Highest resolution shell is shown in parenthesis.

CORRIGENDUM

doi:10.1038/nature16135

Corrigendum: Domains of genome-wide gene expression dysregulation in Down's syndrome

Audrey Letourneau, Federico A. Santoni, Ximena Bonilla, M. Reza Sailani, David Gonzalez, Jop Kind, Claire Chevalier, Robert Thurman, Richard S. Sandstrom, Youssef Hibaoui, Marco Garieri, Konstantin Popadin, Emilie Falconnet, Maryline Gagnebin, Corinne Gehrig, Anne Vannier, Michel Guipponi, Laurent Farinelli, Daniel Robyr, Eugenia Migliavacca, Christelle Borel, Samuel Deutsch, Anis Feki, John A. Stamatoyannopoulos, Yann Herault, Bas van Steensel, Roderic Guigo & Stylianos E. Antonarakis

Nature **508**, 345–350 (2014); doi:10.1038/nature13200

Owing to a labelling error in the input files, one of the two replicate data sets used for Fig. 5d and e and Supplementary Fig. 6d of this Article was incorrect. We have now repeated the analysis with a correct, independent replicate experiment. This confirms our previous conclusion that there are no detectable differences in nuclear lamina interactions between the normal and trisomy 21 twin cells. Therefore, our conclusions remain unaffected. Figure 1 of this Corrigendum shows the corrected panels d and e of Fig. 5. The Supplementary Information to this Corrigendum shows the corrected panel d of Supplementary Fig. 6. The correct data are in the Gene Expression Omnibus under accession number GSE55289.

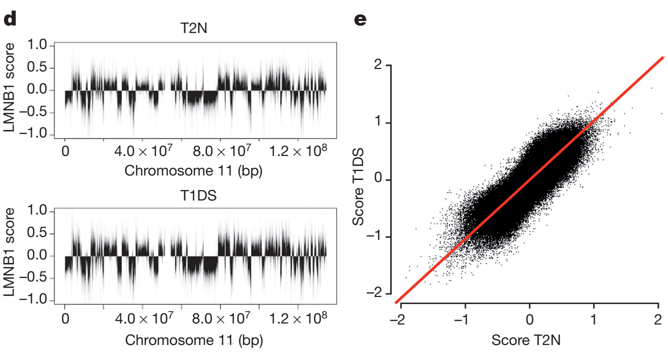


Figure 1 | This shows the corrected panels d and e of the original Fig. 5.

Supplementary Information is available in the online version of the Corrigendum.

CORRIGENDUM

doi:10.1038/nature16470

Corrigendum: Hallmarks of pluripotency

Alejandro De Los Angeles, Francesco Ferrari, Ruibin Xi, Yuko Fujiwara, Nissim Benvenisty, Hongkui Deng, Konrad Hochedlinger, Rudolf Jaenisch, Soohyun Lee, Harry G. Leitch, M. William Lensch, Ernesto Lujan, Duanqing Pei, Janet Rossant, Marius Wernig, Peter J. Park & George Q. Daley

Nature **525,** 469–478 (2015); doi:10.1038/nature15515

In this Review, a sentence was added at proof stages and we inadvertently omitted a citation to a study from the laboratory of Jacob Hanna[1]. This reference citation should have appeared associated with the sentence: "The observation that naive cells tolerate depletion of epigenetic regulators supports the concept of naive pluripotency as a configuration with a reduced requirement for epigenetic repression compared to primed PS cells and somatic cells[1]."

1. Geula, S. *et al.* m^{6}A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* **347,** 1002–1006 (2015).

CORRIGENDUM

doi:10.1038/nature16479

Corrigendum: Failure to replicate the STAP cell phenomenon

Alejandro De Los Angeles, Francesco Ferrari, Yuko Fujiwara, Ronald Mathieu, Soohyun Lee, Semin Lee, Ho-Chou Tu, Samantha Ross, Stephanie Chou, Minh Nguyen, Zhaoting Wu, Thorold W. Theunissen, Benjamin E. Powell, Sumeth Imsoonthornruksa, Jiekai Chen, Marti Borkent, Vladislav Krupalnik, Ernesto Lujan, Marius Wernig, Jacob H. Hanna, Konrad Hochedlinger, Duanqing Pei, Rudolf Jaenisch, Hongkui Deng, Stuart H. Orkin, Peter J. Park & George Q. Daley

Nature **525,** E6–E9 (2015); doi:10.1038/nature15513

During extensive revisions of this BCA, we inadvertently omitted a citation by Takaho A. Endo that used variant calls from RNA-seq data to conclude that the purported Fgf4-induced stem cells (FI-SCs) described in Obokata *et al.*[1] constituted a mixture of trophoblastic and embryonic stem cells[2]. Our analysis, performed independently, reached similar conclusions. We regret this oversight.

1. Obokata, H. *et al.* Bidirectional potential in reprogrammed cells with acquired pluripotency. *Nature* **505,** 676–680 (2014).
2. Endo, T. A. Quality control method for RNA-seq using single nucleotide polymorphism allele frequency. *Genes Cells* **19,** 821–829 (2014).

TECHNOLOGY FEATURE

LIVING FACTORIES OF THE FUTURE

Scientists are designing cells that can manufacture drugs, food and materials — and even act as diagnostic biosensors. But first they must agree on a set of engineering tools.

SPIBER INC.

Japanese company Spiber Inc. has reprogrammed bacteria to make spider silk, which is being used to make clothing.

BY MICHAEL EISENSTEIN

From an evolutionary perspective, yeast has no business producing a pain killer. But by re-engineering the microbe's genome, Christina Smolke at Stanford University in California has made it do precisely that. Smolke and her team turned yeast into a biofactory that, by starting with sugar as a raw ingredient, makes the potent pain-relief drug hydrocodone[1].

This feat is a prime example of synthetic biology, in which scientists reprogram cells to replicate products found in nature — or even make more-specialized materials that would never normally be produced by a natural organism.

Synthetic biologists are ambitious. "We'd all love to imagine a world where we could adapt biology to manufacture any product renewably, quickly and on demand," says Michael Jewett, a synthetic biologist at Northwestern University in Evanston, Illinois. Groups around the world are engineering yeast, bacteria and other cells to make plastics, biofuels, medicines and even textiles, with the goal of creating living factories that are cheaper, simpler and more sustainable than their industrial counterparts. For instance, the biomaterials company Spiber Inc. in Tsuruoka, Japan, has reprogrammed bacteria to churn out spider silk for use in strong, lightweight winter clothing.

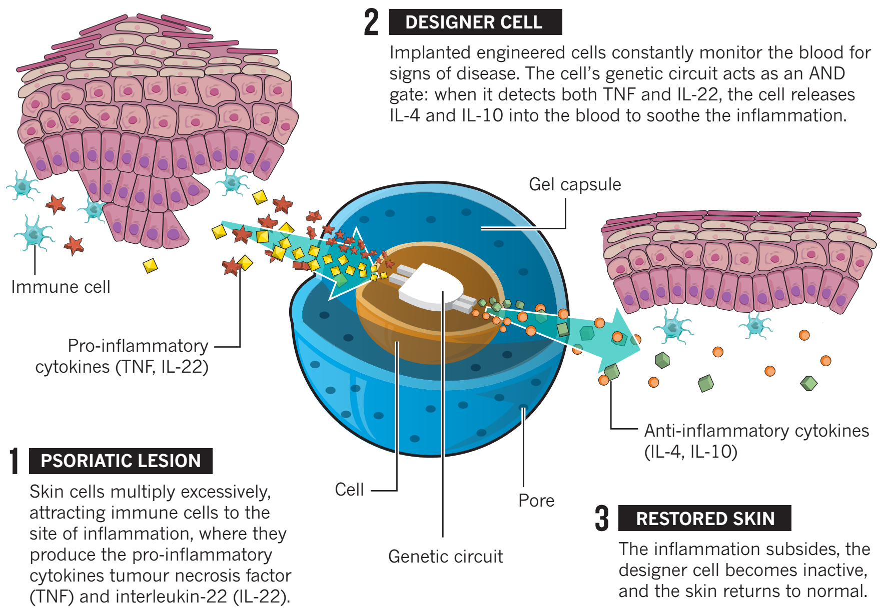
But synthetic biologists are going beyond simply producing materials — they are creating complex systems by 'wiring up' genetic parts into circuits. This approach has already resulted in various living switches and sophisticated sensors. For example, Martin Fussenegger's group at the Swiss Federal Institute of Technology (ETH) in Zurich has built biomedical sensors that can detect disease-relevant metabolites in the blood and trigger the production of therapeutic compounds. In mice, these biosensors successfully staved off gout and obesity, and treated the skin disease psoriasis[2] (see 'Living pills').

This young field has already spawned some success stories, but making and putting together genetic parts currently involves substantial guesswork and unpredictability. For the field to advance, academics and industrial players must agree on a toolbox of reliable genetic parts and the best strategies for assembling them.

To build an artificial product, synthetic biologists begin by selecting DNA parts on a computer and manufacturing them with specialized instruments. The parts can then be inserted into the DNA of microorganisms and cells to reprogram them.

LIVING PILLS

Scientists have made engineered cells that can detect flare-ups of the skin disease psoriasis and dispense on-the-spot treatment. The implanted cells are housed in gel capsules that protect them from the host's immune system and that stop the cells from attacking the host if they malfunction.



ADAPTED FROM REF. 2

Thanks to the plummeting cost of DNA sequencing, there is now a vast collection of genetic data through which synthetic biologists can sift to find useful genes. "Biology has given us this big, crazy library of stuff to choose from," says Christopher Voigt, a synthetic biologist at the Massachusetts Institute of Technology (MIT) in Cambridge. One leading database, the US National Center for Biotechnology Information's GenBank, contains more than 190 million DNA sequences from 100,000 organisms.

Some of the most widely used genetic parts encode enzymes — proteins that are essential for manufacturing. To transform glucose into hydrocodone, for example, Smolke's team took 23 enzyme-encoding genes from diverse species and put them into yeast[1].

Other favourites in the genetic designer's palette are promoters — stretches of DNA that regulate the activity of nearby genes and cause them to be expressed. When proteins called transcription factors bind to a promoter, the process of transcribing a gene begins. But promoters operate too slowly for some synthetic-biology applications. "We're trying to build things that operate fast — on millisecond timescales," says biologist Pamela Silver of Harvard Medical School in Boston, Massachusetts. Scientists are therefore examining alternative mechanisms that allow gene expression to be controlled directly by signals in the environment, such as toxins or antibiotics.

With myriad synthetic DNA pieces at their disposal, synthetic biologists can indulge their creativity. Voigt is enthusiastic about the possibilities: "The nice thing about biology is that there are lots of ways to do the same thing — and as an engineer, you can pick the way that is easiest to design." But genetic parts must perform consistently if the goal of setting up industrial processes is to be realized. "One of the key problems for biology in general is the lack of reproducibility," says Richard Kitney, chairman of the Institute of Systems and Synthetic Biology at Imperial College London. "In synthetic biology this is totally unacceptable — you have to have reproducibility if you're going to do industrial translation."

Many researchers deposit their discoveries into shared repositories, such as the Registry of Standard Biological Parts and the Inventory of Composable Elements. But those parts are often poorly defined or lack crucial information about how they were experimentally tested. "The only quality control is in the person who deposited the information," says Voigt.

The US National Institute of Standards and Technology (NIST) launched the Synthetic Biology Standards Consortium in March 2015 with the aim of standardizing the design, documentation and assembly of synthetic-biology parts across academic institutions and industry. In the United Kingdom, Kitney is coordinating a similar effort in which the DICOM (Digital Imaging and Communications in Medicine) standard for sharing medical information will be expanded to include synthetic biology[3]. In parallel, an international team has developed SBOL (Synthetic Biology Open Language)[4] to provide researchers with a standardized vocabulary to describe genetic parts and circuits.

CELLULAR SOFTWARE

Thanks to greater automation, it is now simpler and cheaper than ever before to make synthetic DNA parts (see 'Manufacturing DNA has never been easier'). But connecting those parts to form genetic circuits that can work together to provide sophisticated, computing-like behaviours is still a challenge. "Any time you physically connect DNA you're creating a new sequence at that interface — as DNA is so information-rich, you could create a new promoter or change the beginning of the RNA," says Voigt.

Even carefully designed circuits can malfunction and cause unwanted expression of a gene or interference between the genetic elements in the biological circuit — outcomes that cannot be foreseen in computer models. "The community is very much operating in a world where we cannot predict what is going to happen in our systems when we build them," says Reshma Shetty, co-founder of the synthetic-biology company Ginkgo Bioworks in Boston, Massachusetts.

This uncertainty means that many of the steps in engineering a synthetic system need to be tested and optimized. Software tools and robotics are speeding up each part of this process, from building the artificial DNA to inserting it into a microbe. "You can use high-throughput prototyping to just build every variant, and hopefully one of them will hit," says Jay Keasling, a biochemical engineer at the University of California, Berkeley, and a pioneer in the field. The push for automation has led a number of synthetic-biology research centres and firms to install 'biofoundry' facilities in which robotic assembly lines create, test and optimize microbes at a much larger scale than could be done by hand.

Biofoundries are enabling synthetic biologists to embark on ambitious projects. For example, Voigt, who co-directs the MIT-Broad Foundry, cites a collaboration with the Swiss pharmaceutical company Novartis to manufacture a huge range of molecules that are produced by bacteria in the human gut.

Other institutions pursuing the biofoundry model include the SynbiCITE programme at Imperial College London and the National University of Singapore's Synthetic Biology Foundry. The US Defense Advanced Research Projects Agency (DARPA) has also invested heavily in the MIT-Broad facility, including a five-year, US$32-million contract that began in October 2015.

Some biologists remain sceptical about the rush to scale up and automate, and favour a more theory-driven strategy. But Kitney, who co-directs SynbiCITE, considers automation to be an inevitable step in the evolution of synthetic biology. "You can rapidly run a whole series of experiments in parallel to see which configuration works best," he says.

THE PERFECT HOST

Species that are commonly used as model organisms in the lab, such as brewer's yeast (*Saccharomyces cerevisiae*) and the bacterium

Escherichia coli, have also been pressed into service by synthetic biologists. Many breakthroughs in biosynthesis have been achieved with these organisms, such as when Keasling and his collaborators at Amyris, a company that he co-founded in Emeryville, California, in 2003, reprogrammed *S. cerevisiae* to manufacture the antimalarial compound artemisinin[5].

But these common lab organisms are not necessarily suited to being grown on an industrial scale. The hunt for better alternatives has led scientists to search in obscure places. "More and more labs are taking on arcane organisms — I think the *S. cerevisiae* and *E. coli* dominance is dropping," says Voigt.

In some cases, the choice organisms will be those that can withstand harsh manufacturing conditions, says Keasling. "Maybe you're producing something that's toxic but volatile, so if you have an organism that can produce it at relatively high temperatures, you could boil it off while you're producing it." Scientists are also testing whether it is possible to feed microbes with carbon sources other than sugars to make products. Synthetic-biology company Intrexon in Germantown, Maryland, is working with bacteria that feed on methane, a cheaper and more efficient means for producing carbon-based products than is sugar.

MEDICAL CELLS

When it comes to medical applications, synthetic biologists are engineering mammalian cells rather than microbes. Such designer cells could produce drugs in response to disease or take over certain physiological tasks in people with metabolic disorders such as diabetes. But engineering mammalian cells introduces a new set of challenges. "All the tools we have in yeast are just not there in mammalian cells," says Smolke. "We don't have as many promoters, or tools for regulation of gene expression or protein modification."

"It will come to the point where you can just inexpensively synthesize the DNA you need."

The easiest cells to cultivate are tumour-like, immortalized cell lines, which are inherently 'defective' and therefore not representative of healthy tissues. Conversely, tissue-derived primary cells are hard to cultivate and manipulate, and differences between cell types confound efforts to build toolkits that can be applied across the body. "Something that works in a kidney cell will not necessarily work in the lung or liver," says Fussenegger. To get around this, the ETH team is engineering 'prosthetic gene circuits', which are introduced into host cells that can be implanted at the site of disease.

Tinkering with genomes can also present problems. Even 'smart' genome-editing tools — such as CRISPR–Cas9, a system for introducing targeted modifications at specific DNA sites — can have unpredictable outcomes. "We don't know enough loci in human cells where you can put things in without any interference," says Fussenegger. His team is exploring whether it is possible to avoid this uncertainty by introducing gene networks that are embedded in synthetic loops of DNA known as plasmids rather than integrated directly into chromosomes. As an extra precaution, his experiments with mice generally make use of engineered cells trapped in implanted capsules, rather than modifying the animal's tissues.

Others want to do away with the cell altogether. Jewett is studying cell-free systems in which bacterial extracts are purified to obtain only the 'useful' parts of the cellular machinery. "You get all the enzymes necessary for energy and cofactor regeneration as well as protein synthesis," says Jewett. "This gives you unprecedented freedom to directly manipulate reaction conditions." This allows researchers to establish chemical conditions that maximize manufacturing productivity without worrying about keeping cells healthy. Jewett's team has shown that this approach can efficiently churn out medically useful proteins such as erythropoietin[6], a hormone that stimulates red-blood-cell production. "It's not yet a replacement for existing technologies, but the yields are sufficient to serve as a complement," he says.

The field is still in its infancy — indeed, the earliest demonstrations of engineered genetic circuits appeared only in early 2000 — and it can be dauntingly complex. Even so, a growing number of scientists grounded in conventional molecular biology are keen to give genetic design a try. Synthetic biologist Ron Weiss at MIT is teaching an online course on the field that proves its popularity. "We've had about 14,000 people sign up," he says.

The pay-off for those entering the field could be huge. "I'm in this space because the frontiers are endless for what biology can do," says Shetty. "It's just a matter of the technology advancing to a point where those new horizons open up." ■

Michael Eisenstein *is a freelance writer based in Philadelphia, Pennsylvania.*

GENES TO ORDER

Manufacturing DNA has never been easier

In the early days of synthetic biology, scientists had BioBricks — a genetic-part format that was conceived by Thomas Knight at the Massachusetts Institute of Technology in Cambridge and designed for modular assembly. It was an attractive concept, but stringing these short bits together to make larger circuits proved to be a laborious and potentially error-prone process.

The task of assembling the pieces is now much easier because newer DNA synthesis machines can churn out strings of several thousand base pairs rather than just a few hundred, which cuts down on the errors introduced by the assembly process. "You can just order a bunch of predictably designed constructs, so you don't even have to think about the modularity anymore," says Pamela Silver of Harvard Medical School in Boston, Massachusetts. The cost of making DNA parts has also fallen dramatically — by as much as 85% between 2009 and 2014 — to the point at which both academic groups and companies routinely outsource the job to specialized providers such as Twist Biosciences in San Francisco, California, Gen9 in Cambridge, Massachusetts, and SGI-DNA in La Jolla, California.

Many projects still require constructs that exceed the scale of what can be manufactured in one go, but these longer fragments can now be joined by quick and simple techniques that leave no scar. However, Jay Keasling of the University of California, Berkeley, thinks that even this will soon become a thing of the past. "It will come to a point where you can just inexpensively synthesize the DNA you need, whether it's 10,000 or a million base pairs." M.E.

ROY KALTSCHMIDT/BERKELEY LAB

Jay Keasling with a Biomek FX[P] lab robot.

1. Galanie, S., Thodey K., Trenchard, I. J., Filsinger Interrante, M. & Smolke, C. D. *Science* **349,** 1095–1100 (2015).
2. Schukur, L., Geering, B., Charpin-El Hamri, G. & Fussenegger, M. *Sci. Transl. Med.* **7,** 318ra201 (2015).
3. Sainz de Murieta, I., Bultelle, M. & Kitney R. I. *ACS Synth. Biol.* http://dx.doi.org/10.1021/acssynbio.5b00222 (2016).
4. Galdzicki, M. *et al. Nature Biotechnol.* **32,** 545–550 (2014).
5. Paddon, C. J. *et al. Nature* **496,** 528–532 (2013).
6. Sullivan, C. J. *et al. Biotechnol. J.* **11,** 238–248 (2016).

RESEARCH DEVELOPMENT Advance cutting-edge science without running a lab **p.406**

RESEARCH TOPICS Do the PhD you want to do **go.nature.com/vgzsjt**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

IZABELA HABUR

Classes and seminars help trainees to get the most out of career-development tools.

CAREER PLANNING

Question time

Career-development plans can point researchers in directions they might not have expected, but they take commitment.

BY PAUL SMAGLIK

In 2014, Michael Burel completed an online workbook that asked about his scientific competencies and interests. When he indicated that he was skilled in statistical analysis and enjoyed presenting research to a non-scientific audience, the programme suggested that he work in public policy, a field that didn't interest him. He tossed the results aside.

But last year, the doctoral student, who is studying stem cells at New York University, revisited the tool as part of a course on building career options. This time it led him to science writing, a path that resonated, and the instructor and guest speakers helped him to identify ways in which he could train for a career in the field.

Since then, he has attended a science-writing seminar, talked to science journalists about how they trained for and landed their jobs and attended a science-writers conference. He now interns as a science writer for the Albert and Mary Lasker Foundation in New York City, which supports medical research. He says that the career workbook and related course have been some of the most useful aspects of his graduate education: together, they helped him to identify a viable career and guided him to the workshops, classes and internships that provided a starting point for his success.

Burel's experience illustrates both the promise and problems of the career-development programmes known as individual development plans (IDPs) in the United States and researcher development frameworks (RDFs) in the United Kingdom and mainland Europe. The programmes, available in hard copy and online, aim to help trainees to identify what aspects of science they like best, match them with careers that incorporate their interests and skills and identify gaps in their competencies.

Versions of the programmes can be as elaborate as the multi-question workbook that Burel initially turned to, or as simple as a short conversation with an adviser followed up by a written training plan. Either way, IDPs and RDFs can lead users in wrong directions and to dead ends when completed on their own or without follow-up. Junior scientists who hope to exploit the value of a career-development plan should complete them as part of a career-building course, discuss them with peers and an adviser and revisit them often (see 'Career planner').

Although IDPs are commonplace in the private sector, the scientific workforce has adopted them fairly recently. Vitae, a UK-based organization that trains and develops researchers around the world, developed an RDF in 2009. In 2013, the US National Institutes of Health (NIH) recommended that principal investigators (PIs) use them with their postdocs and graduate students. The Federation of American Societies for Experimental Biology (FASEB) later created a template for the hard-copy version, and the American Association for the Advancement of Science (AAAS) launched an online version called myIDP. Many institutions have developed their own version.

BREAK OUT OF BIAS

Some trainees say that their institution's IDP programme is written to aim users towards academia. And, they warn, if a PI or adviser is not on board with other career choices, it can be tricky to get effective results from the programme. Gary McDowell, a postdoc at Tufts University in Medford, Massachusetts, completes an IDP every year and discusses it with his PI. He says that both the programme and his PI are academically oriented, so he tries to be realistic about the plan's empirical value. "It's helpful to figure out conferences to go to, papers to plan, skills I need to be developing," he says. "Ultimately, I think any reflection on your career goals, identifying successes in the past year and planning what you need in the next year is helpful, regardless of how you ▶

CAREER PLANNER

How to get started on finding the right path.

If you are a student or a postdoc and your mentor is unwilling or unable to help you with an individual development plan (IDP) or researcher development framework (RDF), here are some tips to help you get started and to follow through.

- Use an online tool such as myIDP (http://myidp.sciencecareers.org) or RDF Planner (https://rdfplanner.vitae.ac.uk). These programmes work best with follow-up from a mentor, but the software can still help to identify scientific strengths and professional preferences and suggest possible careers, and it offers articles on how to train for them.
- Meet regularly with peers to collectively discuss your IDPs. Peers can often identify and offer suggestions about each other's interests and transferable skills and may be able to point to training resources.
- Develop and tap into a broader network. This can help when students and postdocs suspect that their supervisor or mentor might be biased towards an academic career path. Attend local and regional meetings of professional organizations in your desired career area. Find alumni from your institution who work in the field (the university alumni office can help), and contact them.
- Commit to your plan. An IDP is useless if it sits on a shelf or in your digital device. Take the steps it identifies — building skills and expanding training through courses, seminars and workshops — to fill in gaps.
- Return to your plan regularly. Check on your progress and update any skills or experiences you've gained that could help in your career search. Set new goals and create deadlines for them. P.S.

feel about your career path. But people may fall through the cracks."

There is little formal incentive to complete an IDP or RDF. The NIH does not follow up on its recommendation, and not all institutions require trainees to use one. Nor is there a mandate for its use in any nation. Just 47% of the postdoctoral offices that participated in a survey by the US National Postdoctoral Association said that they require their postdocs to complete an IDP, according to a 2014 report (see go.nature.com/awsupm). And another 37% encourage their use, the report said.

Ultimately, the trainee should not just create, but also follow through on their career-development plan, says Philip Clifford, associate dean for research at the University of Illinois at Chicago, who helped to develop both the FASEB and AAAS versions. The biggest mistake users make is to consider it an endpoint rather than a launch pad, he adds. He has run some 200 career-building courses and workshops that incorporate the plan and build on its use and results.

Lina Dimberg, who participated in one of Clifford's courses, followed the instructions with care and found the programme fruitful. As a postdoc in cancer research at the University of Colorado Denver, she knew that she did not want to stay in academia but was unsure of her options. She completed an IDP with Clifford's guidance and immediately learned that her strongest skills — writing grant proposals and papers, reading the literature and discussing research — gave her a solid foundation for several science-related careers, one of which was medical writing.

Clifford's curriculum required her to set up meetings with scientists in occupations that the IDP had pinpointed as career possibilities. One of those chats led to a job as a writer at a medical-device company after her postdoc ended; today, she works there as a senior scientist. Dimberg credits the IDP process and the workshop that supported it for helping her to define her career objectives and to develop the necessary confidence to market herself for the position. "The IDP opens your eyes to careers where you can combine your science interest with other interests and skills," she says.

ALL IN THE TIMING

Sometimes, the programme might identify a good direction, but the user might not be quite ready at that point in time. When Nathan Vanderford was a graduate student and postdoc, his PIs encouraged him to complete an IDP that highlighted a tenure-track position — even though he wasn't sure that was the right route. "I ended up with a plan that I felt was not true to my desired career path," he says. He then spent years in other careers, including science communications and research operations.

Today, he has come almost full circle and is now a faculty member at the University of Kentucky in Lexington, where he has a faculty-administrator post that his IDP results from so long ago did not quite predict. He teaches a career-development class that incorporates the programme's best principles. "The IDP I was forced to do has little to do with my current position," he says. "I want to give students a mechanism that allows them to explore freely any career option they want to pursue." ■

Paul Smaglik *is a freelance writer in Milwaukee, Wisconsin.*

TRADE TALK
Funding fixer

MONICA HOLDER

After a PhD and postdoc in cardiovascular biology, Christina Papke moved into research development. She works at Texas A&M University in College Station, where she assists biomedical faculty members with grant proposals and helps them to form collaborations and identify funding opportunities.

When did you start exploring careers beyond running a lab?

The first and most difficult step was to realize that I did not want to be a principal investigator. In summer 2014, a series of both difficult and good events — a grandparent's death, a friend's wedding, a crime in my apartment complex — left me feeling like a rubber bouncy ball emotionally. It caused me to think, pray — and re-evaluate the direction of my career. I realized that although I enjoyed thinking about science, I could not see myself working in a lab for the next 30-plus years.

How did you learn about your current job?

What was extremely helpful for me was joining the American Medical Writers Association (AMWA) and being willing to put myself out there at a conference, not knowing who I might meet. Two people presented on being a grant-writing professional. That required a solid research background, as well as understanding how grants work, communicating with scientists, coordinating events and being an information resource. That was the combination I didn't know I was looking for. I realized that I could take my favourite aspects of research with me.

So the session clinched it for you?

I wouldn't have found this job if I hadn't talked to the presenters. One of them said, "I know a place." It was a city I hadn't been looking in. I joined AMWA in February 2015, went to the conference in April and was hired in July.

That sounds almost preordained.

One of my mentors said to me, it's like putting out a lot of fishing lines. I was also doing informational interviews and a variety of other things. But during the process, it looked quite messy. My advice is, sometimes you don't know what you're looking for. And that's okay. ■

INTERVIEW BY MONYA BAKER

This interview has been edited for length and clarity. See go.nature.com/kbvz2q for more.

GENIUS LOCI

Interactive fieldwork.

BY S. R. ALGERNON

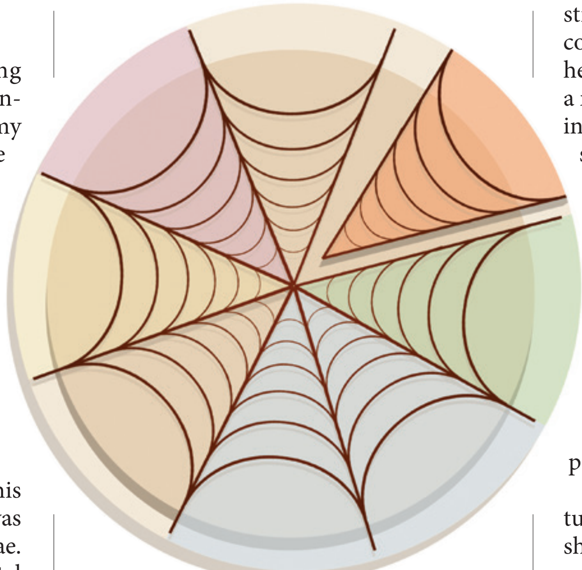


ILLUSTRATION BY JACEY

At first, I mistook the skittering footfalls outside my tent for thunder. I jammed my pillow over my ears and returned my attention to the geosurvey data on my phone. After two weeks on this backwater planet, I didn't have even half the data I needed. I had timed my visit to coincide with an alignment of the planet's three moons. According to Dr Feldman, that would set the atmosphere abuzz and be the perfect test of my hypothesis that the crystalline formations on the planet's surface were a product of the planet's mesmerizing storms.

Nobody told me that, for the locals, this was Synergy Month. The whole village was full of clanging and the squealing of pupae. The weather had cut me off from the orbital communications hub, so I couldn't even watch through the orbiting satellites. Worst of all, the locals who were supposed to help me carry equipment up to the mountains turned me away, even when I doubled the price. *We cannot*, they said. *The Spirit-Storm gathers. The Queenling beckons.*

Explain that to my thesis committee.

With only a few more days on-planet, I found a tent and set off for the mountains myself. I couldn't carry a full complement of scanners, but I could make do with the apps on my phone and hope to clean up the data afterwards.

Outside, lightning struck close enough to cast shadows of exoskeletons and segmented limbs against the walls of my tent. A spiny leg tore through its fabric and impaled a corner of my sleeping bag.

I wriggled free, stuffed my phone into a trash bag and managed an insectoid version of running with the bulls until I lunged onto the rocks by the roadside.

That was close. I recalled a quip from Dr Soto's Xeno-archaeology class. *Out on the frontier, the bugs step on you.*

The insectoid that had ruined my tent extricated its leg and weaved through the crowd towards me.

"Apologies," it said. "I hope I did not injure you."

"Just a scratch or two."

"You should join the procession. The Spirit-Catcher will find a spirit that will heal your wounds."

➲ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

"I'd rather take my chances with the clinic at the spaceport," I said, as I dusted myself off.

"Spaceports come and go. The Spirits will always be with us," said the insectoid pilgrim as it rejoined the procession. "Have a safe journey."

I climbed up the hillside and watched the storm. Now free from the distracting footfalls and echoes down below, I noticed that most of the lightning arced towards a ridge in the distance. I trudged to the top of the ridge, and it turned out to be the weathered edge of an impact crater. Lightning had scorched and melted the rock, leaving it alternately pockmarked and slick.

I photographed the crater surface. Three more strikes etched a radiating network of lines onto my vision through my closed eyelids. My skin tingled. My ears rang. The lightning pooled or collected somehow in the crater. *Why would lightning seek low ground? Was something catching it?*

Maybe I could get a dissertation out of that Spirit-Catcher nonsense after all.

I aimed my phone at the glowing spot in my field of vision and stepped forwards. The slope was steeper than I thought, but I held my ground. I felt eerily calm. All my training told me to get clear, but the crater drew the energy to its centre. After recording a dozen more strikes, I turned to walk back to the trail.

The mud slipped underfoot. I fell to one knee and pushed off with my hands. As I regained my footing, strands like spider silk clung to my palms and fingertips.

Lightning flashed once more. The strands glowed. Every muscle in my body convulsed and I tumbled backwards. As my head struck the ground, pain sent me into a nauseated daze. As I fell, fibres gathered into a cocoon around me — a shimmering, silvered Faraday cage.

I awoke at the bottom of the crater, hanging from a web by strands that spanned the crater floor. They supported my weight like marionette strings. The cocoon enveloped each leg, arm and finger individually, so it did not encumber my movement.

Lightning struck. The strands of the thick web above me glowed as the current passed overhead.

A rustling sound caught my ear, and I turned to see a twelve-legged arachnoid shape emerge from an alcove in the shadows. As it approached, my arms pulled tight against my chest in a dead-Pharaoh pose. The strands connecting me to the web pulled taut, lifting me off the ground.

"You must be the Spirit-Catcher," I said, grateful that the cocoon permitted me to speak.

"When the others arrive," said the Spirit-Catcher, "do as I instruct you and do not say a word." The cocoon tightened around my neck and then released. "I will know if you try to deceive me."

The Spirit-Catcher inspected the corners of the web, which held hundreds of battered machines and dozens of non-human skeletons.

The web that confined me disgorged a dented, blood-flecked medi-bot, which rolled towards me.

"You can use this, yes?"

I nodded. All the field researchers had to learn as part of their training.

The Spirit-Catcher seemed satisfied and scuttled back to the alcove. The lightning stopped. I heard a regal voice echo through the chamber.

"All praise the Spirit-Catcher. Bring forth the injured and infirm."

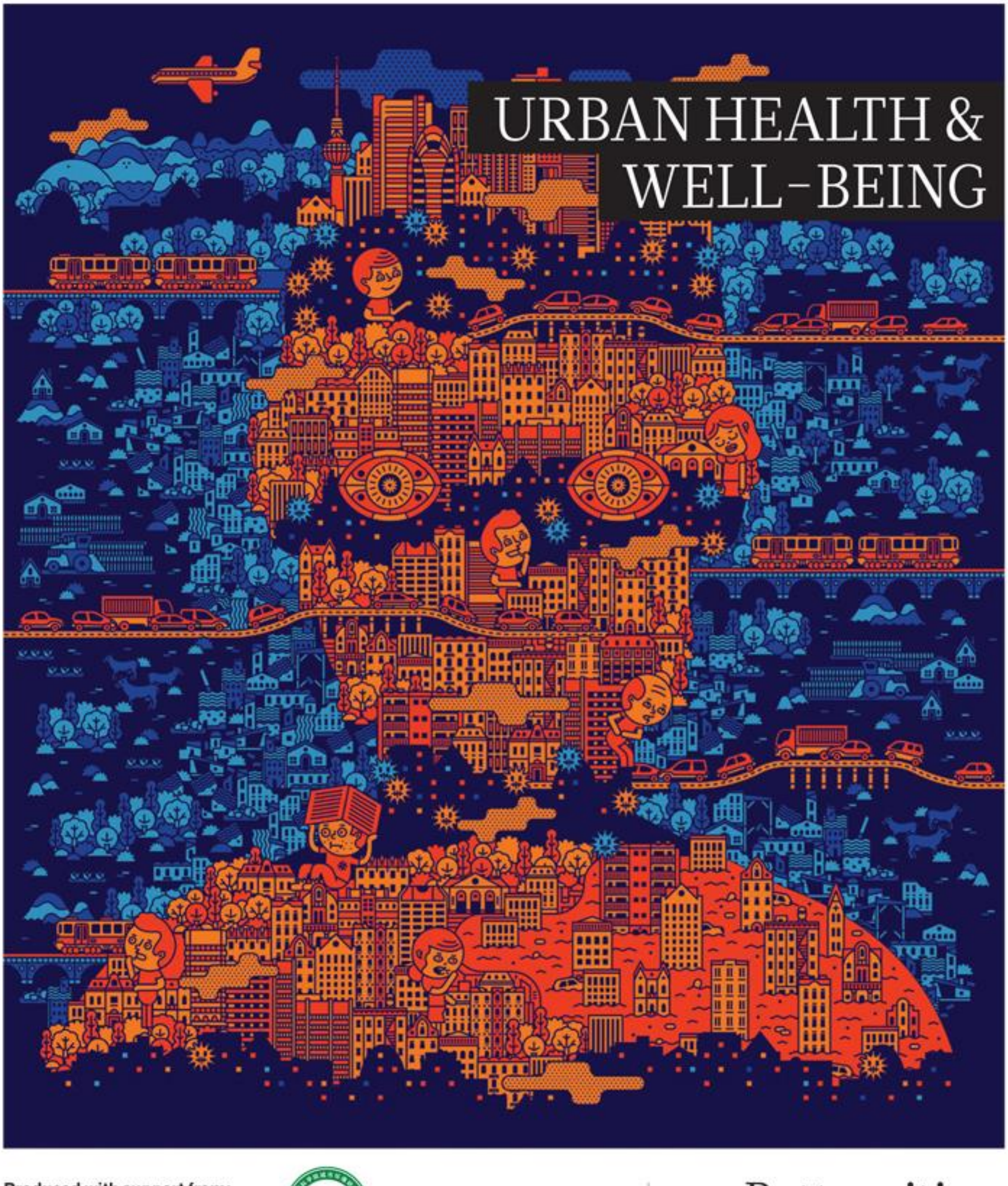
I told myself they would let me go once the procession had left, but the pilgrim's words stuck with me.

"Spaceports come and go. The Spirits will always be with us." ■

S. R. Algernon *studied fiction writing and biology, among other things, at the University of North Carolina at Chapel Hill. He currently lives in Singapore.*

nature OUTLOOK

17 March 2016
Supplement to Nature Publishing Group journals



URBAN HEALTH & WELL-BEING

Produced with support from:

中科鼎實環境工程有限公司
ZHONGKE DINGSHI ENVIRONMENTAL ENGINEERING CO.,LTD

IUE

URBAN HEALTH AND WELLBEING
A SYSTEMS APPROACH

汇绿园林建设发展有限公司
HUILV LANDSCAPE CONSTRUCTION DEVELOPMENT CO.,LTD

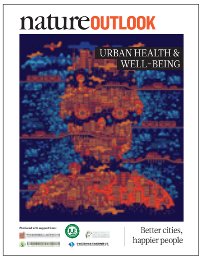
宁波天河水生态科技股份有限公司
Ningbo Tianhe Aquatic Ecosystems Engineering Co., Ltd.

Better cities, happier people

nature OUTLOOK

URBAN HEALTH AND WELL-BEING

17 March 2016 / Vol 531 / Issue No 7594



Cover art: Jan Kallwejt

Editorial
Herb Brody,
Michelle Grayson,
Richard Hodson,
Jenny Rooke

Art & Design
Wesley Fernandes,
Mohamed Ashour,
Andrea Duffy

Production
Karl Smart,
Ian Pope,
Matthew Carey

Sponsorship
Stella Yan,
Samantha Morley,
Stella Jiang

Marketing
Hannah Phipps

Project Manager
Anastasia Panoutsou

Art Director
Kelly Buckheit Krause

Publisher
Richard Hughes

Chief Magazine Editor
Rosie Mestel

Editor-in-Chief
Philip Campbell

The world is in the midst of the largest wave of urban growth in history. With more than half the population already living in cities, and further rises expected, the health of city dwellers is crucial to global well-being.

Throughout history, many of the innovations designed to improve city living have instead posed a challenge to having a healthy, happy life (page S50). Urban environments designed around the car, for instance, promote a sedentary lifestyle (S52).

Faced with shifting climates, urban policymakers are preparing to deal with an increased risk of flooding (S54) — the replacement of concrete with more absorbent surfaces may be part of the solution. Evidence is also building that green space, although scarce, is a valuable resource for improving mental health — especially in poorer communities (S56).

Health inequality is the bane of many countries. A community project in California shows how tackling stress in deprived neighbourhoods can improve the health of people whose life expectancy often lags behind that of their richer neighbours (S58). Although improved living standards have reduced the toll of infectious disease over the years, the world's poorest urbanites remain vulnerable. From cholera to Zika virus, slums provide the ideal place for pathogens to thrive (S61).

Interventions to enhance urban living, such as city farming, are not without their pitfalls (S60). Although navigating the risks of urban improvement will not be easy, it is essential if we are to better the lives of the urban population (S64).

We are pleased to acknowledge the financial support of the Institute of Urban Environment, Chinese Academy of Sciences; Zhongke DingShi Environmental Engineering Co., Ltd; Ningbo Tianhe Aquatic Ecosystems Engineering Co., Ltd; and Huilv Landscape Construction Development Co., Ltd in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

Richard Hodson
Supplements editor

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, *Nature* **Vol. XXX**, No. XXXX Suppl., Sxx–Sxx (2016).

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Urban Health and Well-being* supplement can be found at http://www.nature.com/nature/outlook/urban-health
It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
For UK/Europe: Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group — Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

CUSTOMER SERVICES
Feedback@nature.com

CONTENTS

The rise of the urbanite

~10,000 BC

Neolithic people in Mesopotamia, now roughly Iraq, Kuwait and Syria, swap nomadic hunter–gatherer lifestyles for life in villages, near which they grow crops and keep livestock. These settlements attract vermin, insects and parasites, and the denser populations that they support encourage infectious diseases to emerge. Many animal pathogens, including those that cause tuberculosis and smallpox, make the jump to humans.

~6000 BC

Recognizable cities appear in Mesopotamia (**pictured**) on the floodplains of the rivers Tigris and Euphrates. Other cities follow: in the Indus Valley, now Pakistan; on the banks of the River Nile in Egypt; and in the east along the Yellow River in China. As they grow, so does the production of waste. Contaminated and stagnant water trigger outbreaks of disease.

~3200 BC

Shahr-e Sūkhté — the Burnt City — in southeast Iran installs a system of pipes to supply clean water as well as sewers to discharge waste. In the Indus Valley, the grid-based cities of Harappa and Mohenjo Daro manage waste with drains in the streets and brick-lined sewers (**pictured**). Most houses have a private well and toilet.

100

The population of Rome reaches 1 million. According to the Greek geographer Strabo, while the Greeks build beautiful cities, the Romans focus on "paving their roads, constructing aqueducts, and sewers." Aqueducts provide more than 1,000 litres of water per person, per day — far beyond the amount that people use today. Yet hygiene is poor, diseases are rife and mortality is high.

1348

Bubonic plague — the Black Death — sweeps across Europe. The pandemic prompts many city authorities to restrict the movement of residents. In Italy, the cities of Venice, Florence and Lucca set up boards that have the power to impose — and enforce — quarantine.

ANCIENT METROPOLIS ~ 2600 BC

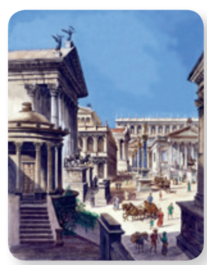
The city of Caral takes shape in the Supe Valley, Peru on the slopes of the Andes mountain range. The first major city in the Western Hemisphere, Caral boasts massive pyramids and extensive residential complexes that cover an area of 65 hectares.

FLUSHED WITH PRIDE ~ 600 BC

Tarquinius Priscus, the King of Rome, orders the construction of the Cloaca Maxima (**pictured**), or the Greatest Sewer, to drain the marshes that surround the city and to flush waste into the River Tiber. Over the centuries, the sewer is expanded, and it is still in use.

DIN CITY ~ 100–110

The Roman satirical poet Juvenal writes of the "thousand perils" that are faced by those who live in ancient Rome (**pictured**). These include fire, collapsing buildings — and insomnia caused by night-time traffic. He states that the noise from passing wagons would "rouse a dozing seal — or an emperor."



FIERY WORDS 1661

English diarist John Evelyn presents King Charles II with *Fumifugium* — a discussion of air pollution in London, in which he likens the city to the "suburbs of hell". Evelyn suggests solutions such as a switch to cleaner fuels.

CLOCKWISE FROM TOP LEFT: PRIVATE COLLECTION/BRIDGEMAN IMAGES; LISZT COLLECTION/BRIDGEMAN IMAGES; LOOK AND LEARN/BRIDGEMAN IMAGES; HULTON ARCHIVE/GETTY IMAGES; DEA/G. NIMATALLAH/GETTY IMAGES

From their earliest beginnings, cities have brought both benefits and risks to the health of their inhabitants. Although some of the hazards have been banished, others remain — and new ones have emerged. By **Stephanie Pain**

~1750

The Industrial Revolution triggers the first large wave of urbanization. Across Europe, rural dwellers flock to the cities that are developing around centres of manufacturing. The population of London soon reaches 1 million and continues to grow as more migrants come in search of work. Cramped housing and poor sanitation lead to epidemics and high mortality.

1854

Cholera rages through London's Soho district, and the outbreak is traced to a well that is contaminated by sewage. London's sewage problem persists until the summer of 1858 when the stench of the River Thames reaches Members of Parliament in the Palace of Westminster. A network of sewers and pumps designed by visionary engineer Joseph Bazalgette (**pictured**) is hurriedly approved. After its completion in 1875, London never experiences another cholera epidemic.

1893

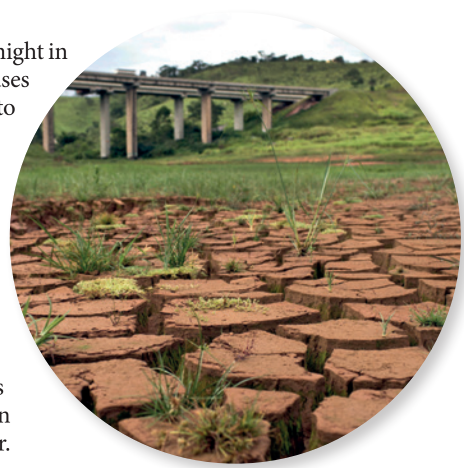
People converge on the World's Columbian Exposition in Chicago, Illinois, to admire the White City. The architectural exhibit embodies the 'city beautiful' concept, which proposes that decaying urban centres are replaced with classical architecture, parks and lakes to remedy social unrest and crime. In the United Kingdom, the garden city movement also advocates green spaces, fresh air and walking.

1943

A thick haze that irritates the lungs settles over wartime Los Angeles, California. The cause remains unknown until Arie Haagen-Smit, a chemist at the California Institute of Technology in Pasadena, solves the mystery — the haze consists mainly of ozone, which is formed by the action of sunlight on the emissions that spew from car exhausts.

2003

In February, a doctor from Guangdong in China spends a night in a hotel in Hong Kong. He is infected with the virus that causes severe acute respiratory syndrome (SARS), which spreads to other guests who then carry it around the world. By June, almost 8,500 people in 30 countries have contracted SARS.



2015

Brazil is experiencing its worst drought for 80 years. São Paulo, a megacity of almost 21 million people, starts to ration water as reservoirs dry up — and blackouts occur as hydroelectric power stations shut down. Water hoarding in open tanks brings mosquitoes and with them dengue fever.

MASS MOVEMENT 1863

London opens the world's first underground railway line (**pictured**), which uses coal-fired steam trains. The city soon has a network of subterranean electric lines and trains that form part of the earliest mass-transport system.

UPWARDLY MOBILE 1885

US architect William Le Baron Jenney builds the first skyscraper in the world, the ten-storey Home Insurance Building (**pictured**) in Chicago. Its metal frame allows tall structures to be built without the need for thicker walls — an innovation that enables high-density city living.

GREEN SHOOTS 1992

The UN Earth Summit in Rio de Janeiro, Brazil, pushes sustainability and climate change up the political agenda — which prompts leaders to take action. Planting trees and creating green spaces are proposed as ways to soak up excess rainwater and to mop up pollutants.

DEAD HEAT 2003

Europe experiences its hottest summer for 500 years, which kills an estimated 70,000 people. Those in cities are the most affected as a result of an urban heat-island effect, and morgues run out of space as the death toll rises.

TIPPING POINT 2008

More than half of the world's population now lives in urban areas. In more-developed nations, the figure is around 74%; in less-developed ones, 44% live in cities. By 2100, 85% of the world's population is expected to be urban.

CLOCKWISE FROM TOP LEFT: LOOK AND LEARN/BRIDGEMAN IMAGES; THE PRINT COLLECTOR/ GETTY IMAGES; ROOSEVELT CASSIO/REUTERS; BETTMANN/CORBIS; CORBIS

JAN KALLWEJT

MOBILITY

The urban downshift

Transporting people around the cities of the future is a public-policy challenge, but it's also an opportunity to improve the health of urban populations.

BY SARAH DEWEERDT

Many cities around the world are spreading faster than their populations are growing. According to researchers at New York University, between 1985 and 2000 the population of Accra in Ghana increased by 50%, but its land area grew by 153%. People are having to travel further: in Nairobi, for example, the average commuting distance increased from less than 1 kilometre in 1970 to 25 kilometres in 1998. As this trend continues, governments face the problem of how to move people around ever-expanding metropolises efficiently enough that residents can take advantage of the opportunities — economic and otherwise — that cities have to offer.

This issue is certainly a public-policy challenge, but it's also an opportunity to improve the health of the world's growing urban population. And researchers and public-health experts say that making cities of the future function well and support human health may depend on the most low-tech, ancient assets available — our own two feet.

"If the pedestrian is happy and you see many pedestrians, that's a city with a good transport system," says Clayton Lane of the Institute for Transportation and Development Policy, a non-profit organization based in New York. "The pedestrian is the indicator species for a sustainable transport system" — and, as it turns out, for a healthy one.

But getting there will require a major shift in government spending priorities and in public attitude. "In many cities around the world, the people and the politicians have this vision of modernity that prominently features automobiles," Lane says. As residents become wealthier, the urban infrastructure is remodelled to favour cars. Many cities in the developing world spend around 70% of their transportation budgets on car-oriented facilities, even though around 70% of trips take place on foot or by public transport, Lane says. The result is that the world is on track to have 2.3 billion cars by 2050. That's just over double the number that were on the road in 2010, and it represents a major threat to the health of the urban population.

ROAD TO HEALTH

Cars promote a sedentary lifestyle, with its attendant risk of health problems such as obesity and heart disease. Driving, especially in congested traffic, causes stress, and air pollution worsens respiratory diseases such as asthma.

According to the World Health Organization (WHO), 1.3 million people die from traffic accidents and 3.2 million from lack of physical activity every year. Outdoor air pollution causes 3.7 million deaths annually, and land traffic is responsible for around 5% of deaths caused by fine particulate matter and ozone[1]. Cars are responsible for only a portion of that pollution (two-stroke scooters are thought to contribute a disproportionate amount to

urban air pollution), but are "the worst mode of transportation in terms of all these categories," says Jeff Speck, an urban planner and author of *Walkable City* (Farrar, Straus and Giroux, 2012).

Much better is what researchers call active transport: walking and cycling, mainly. Sitting in traffic is stressful, whereas physical activity boosts happiness (see page S56). Fewer cars on the road means less choking pollution and fewer deaths in car crashes. Furthermore, the more that people walk and cycle in the city, the safer these activities become — both because there is safety in numbers and because cities provide infrastructure to accommodate these activities.

In general, people are healthier when they are able to do most of their day-to-day activities and errands on foot. For example, rates of childhood obesity are lower in more walkable neighbourhoods[2]. Among older adults in low income neighbourhoods, people living in walkable areas have a lower body mass index than those living in areas where moving around on foot is difficult[3]. And moving from a low-walkability to a high-walkability neighbourhood decreases the risk of having high blood pressure[4]. "Most cities uniformly seek to improve mobility for their citizens, and I think that itself is worth questioning," says Speck. "Mobility is often seen as the ideal, when in fact what we really want is access." That means urban planning needs to emphasize not just moving people around efficiently, but also making sure people's needs can be met nearby.

Of course, walking and cycling are not by themselves sufficient to meet people's transportation needs, especially in the growing number of megacities (those with 10 million residents or more). But walking and public transport support each other. A walkable city needs good transport to move people around. By the same token, walkable neighbourhoods make transport systems more cost-efficient to build and help to ensure that they are well used. Some studies have found that access to public transport improves physical activity and health, largely because it gets people walking.

DESIGN FLAWS

A well-designed city can encourage habits that promote good physical health. "Walking is a very simple physical activity that most people can do," says Yan Kestens, who studies how the built environment contributes to public health at the University of Montreal, Canada. But making an environment more walkable can be challenging — especially in cities that took shape after the advent of the car. "The physical structure of our cities lasts for centuries," says Lane. "If we build our cities and suburbs for cars, it's very difficult to retrofit them for walking."

According to British geographer Adam Davies, who recently collaborated with researchers at Yahoo Labs on an analysis of 7-million geotagged photos taken in central London, walkability is hugely compromised by street networks designed around the car. "The more cars and the more lanes of traffic, the less human-friendly that particular street probably is," he says.

There are many reasons that people choose not to walk to destinations that are within walking distance, Davies says. Lack of pavements, inconveniently placed pedestrian crossings, and the need to cross a major thoroughfare, for example, can make walking unappealing or even unsafe. And for some older people, or anyone who has trouble walking, factors such as these can erode walkability surprisingly fast. "I've heard of stories where people take a taxi to go across the street," says Verena Menec, a healthy-ageing researcher at the University of Manitoba in Winnipeg.

Unpicking the subtle barriers that drive people to this kind of extreme is tough — not least because people's real-world behaviour is difficult to predict. Menec and her team[5] asked middle-aged and older adults about their walking behaviour and attitude to walkability. Nearly 60% said it was important to have a grocery store within walking distance. But of this group, 76% said that they drive there rather than walk. "If people are still within their 'car mode' they will probably not actually walk," Menec says.

"Mobility is often seen as the ideal, when in fact what we really want is access."

But some studies have failed to show clear links between walkability and better health. One analysis[6] of data from the Nurses' Health Study — a large, long-term epidemiological study of women in the United States — for instance, found that women living in walkable neighbourhoods are exposed to higher levels of harmful air pollution. But this relationship varies in different parts of the country, suggesting that exposure to pollution isn't inevitable. And, the more people that walk rather than drive, the cleaner the air will be.

A similarly puzzling set of results comes from one of Kestens' studies[7], which found that low-income residents in Montreal were less likely to walk to places than wealthier people, even in parts of the city that were relatively good for walking. Kestens says that the field needs more qualitative studies to uncover the reasons why people do or do not choose to walk. He is also using GPS and wearable devices to more precisely measure how people get around their cities. The goal is to use those insights to help design more walkable urban landscapes.

FIRST STEPS

Communities and urban planners around the world are coming up with creative ways to improve neighbourhood walkability. A grassroots effort in the Indian city of Chennai, for example, is addressing conditions faced by many of the world's poorest urban dwellers. They have no choice but to get around on foot; however, they do so on streets that are not particularly good for walking. "We have city after city where many people are walking, yet the city is not walkable at all," Lane says. In such environments, pedestrians are especially vulnerable to injury and death from traffic accidents, according to the WHO.

The Chennai government has committed to spending at least 60% of the city's transportation budget on measures to encourage walking and cycling. By 2018, the city is aiming to make 80% of its roadways 'complete streets' — wide pavements, bike lanes, space for public transport and organized parking, as well as lanes for cars.

Another piece of the puzzle is developing public transport systems to link walkable neighbourhoods that are within the reach of cities, using scarce financial resources. Bus rapid transit (BRT) has emerged as a practical, affordable solution for many cities, says Lane, whose organization wrote a set of BRT standards, because BRT lines are much faster and cheaper to build than rail-based systems. Yet, they are fast and efficient — they have dedicated lanes, preferential treatment at intersections, and platforms to help people board faster.

Curitiba in Brazil built the world's first BRT network in the 1970s, with the intention of concentrating urban development around bus stops along the route — a planning tactic known as transit-oriented development. Although successful for a time, the city's rapid growth eventually overwhelmed the capacity of their plan. Curitiba now intends to revisit the strategy with a new BRT line and an associated development corridor.

One of the latest converts to the BRT approach is sprawling Accra, which is building a line between the suburb of Amasaman and the city centre. Until now, the city's transport system has been dominated by licensed minibuses known as tro tros, but these only service a little over half the routes that they are licensed to run on. Accra is like many cities in the developing world that lack a functioning mass transport system. "There's such a huge, huge deficit to address, but BRT is a good solution to do it quickly and affordably," Lane says. "High-quality transit is key to a walkable city so you can access other parts of the city that are also walkable." ■

Sarah DeWeerdt *is a freelance science writer in Seattle, Washington.*

1. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. *Nature* **525,** 367–371 (2015).
2. Saelens, B. E. *et al. Am. J. Prev. Med.* **42,** e57–e64 (2012).
3. Van Cauwenberg, J., Van Holle, V., De Bourdeaudhuij, I., Van Dyck, D. & Deforche, B. *Health Place.* **37,** 16–26 (2016).
4. Chiu, M. *et al. Environ. Health Perspect.* http://dx.doi.org/10.1289/ehp.1510425 (2015).
5. Menec, V. H., Brown, C. L., Newall, N. E. & Nowicki, S. *J. Aging Health* http://dx.doi.org/10.1177/0898264315597352 (2015).
6. James, P., Hart, J. E. & Laden, F. *Environ. Res.* **142,** 703–711 (2015).
7. Steinmetz-Wood, M. & Kestens, Y. *Prev. Med.* **81,** 262–267 (2015).

AP/PRESS ASSOCIATION IMAGES

Flooding at a station in Koshigaya near Tokyo.

FLOODING

Water potential

Water is a necessity for any city, but too much of it can threaten lives and infrastructure. As climate change looms, new approaches can help to turn a threat into a resource.

BY JAMES M. GAINES

Water is vital, but as anyone who has felt the effects of flooding will attest, it is possible to have too much of a good thing. For many nations, flooding has always been a problem. Take the Netherlands, for instance: 26% of the country — and 21% of its population — is below sea level, and only a complex system of dykes, pumps and dams, constructed over hundreds of years, keep it dry. But in other countries, flooding is a more modern and growing concern — especially as absorbent rural land continues to be subsumed by the impermeable concrete of our rapidly expanding cities. Between 2008 and 2010, 62% of Chinese cities were flooded. And 39% of 351 cities in the country reported at least 3 serious floods in that time.

In 2012, the heaviest rainfall in more than 60 years killed 79 people in Beijing. Climate change is likely to make such events more common. According to researchers at the Massachusetts Institute of Technology in Cambridge and Princeton University in New Jersey, hurricanes of magnitudes previously seen only about once a century could hit land every 3–20 years by the end of the century[1]. And, in 2013, researchers estimated that the global financial impact of flooding in coastal cities would increase from US$6 billion per year in 2005 to $52 billion by 2050 (ref. 2).

The threat climate change poses isn't so much new as it is an amplification of existing flooding hazards; those already vulnerable will find themselves more at risk. And with cities' high population density and often coastal locations — three-quarters of large cities are found on the coast — they represent a great accumulation of risk. More than half of the world's population already lives in cities and towns, and with the global urban population set to reach 5 billion by 2030, a single flood has the potential to affect millions of people's lives.

How climate change is likely to affect the risks posed to cities by flooding is a question that researchers, developers and city officials around the world are working to understand. Faced with the certainty of some degree of change, a few cities are already developing infrastructure to allow them to control excess water — and potentially even turn it into a resource.

RISING TIDES

Kristina Hill, an urban designer at the University of California, Berkeley, is more familiar than most with the impact that flooding can have on an urban area. She worked in New Orleans as part of the Dutch Dialogues — a Dutch–American collaboration that brought together engineers, architects and designers after New Orleans was ravaged by hurricane-induced flooding in August 2005 to ensure that the city would be better prepared in the future.

Hill is applying what she learned in New Orleans to the San Francisco Bay Area. Many parts of California are experiencing a historic drought, but she believes that in the near future coastal cities such as San Francisco will have to deal with flooding, caused not by falling rain, but by rising tides.

Sea level in the Bay Area is projected to rise by around 1 metre by the end of the century. This has the potential to overwhelm defences and inundate beach-side areas, but Hill is concerned about another, subtler effect of sea-level rise — water bubbling up from underground.

"One of the things most people have forgotten, which the Dutch pointed out to me," she says, "is the groundwater connection to seawater rise."

Seawater is constantly seeping inland, deep underground. Because of its greater density, the salty water tends to push underneath freshwater aquifers. The effect of this is that, when sea levels rise, the level of the fresh groundwater also moves up.

This higher water table is likely to exacerbate flooding from heavy rains, Hill explains. Just as an already sodden sponge is less effective at mopping up liquid, so the ground loses some of its ability to soak up water from the surface. But even without a storm, if sea levels rise high enough, the fresh groundwater will be pushed to the point at which it begins to leak into underground structures, or up through the surface. "We'll end up with lots of flooding driven by groundwater," says Hill.

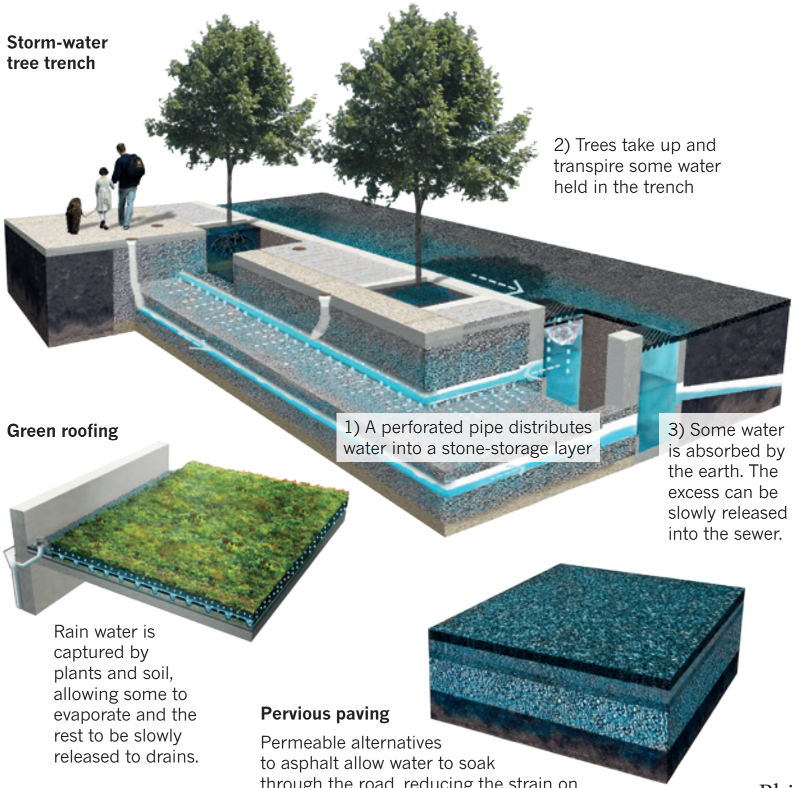
The effect that sea-level rise could have on groundwater in the Hawaiian city of Honolulu could more than double the amount of flooding that the city will experience owing to inundation from the sea alone, researchers have suggested[3]. And, in Miami, Florida, people have reported groundwater bubbling up through their gardens at high tide.

For Hill, the solution lies not in preventing sea-level rise, but in adapting to it. "We have to learn to live with our feet wet," she says. Waterproofing infrastructure such as the subterranean tunnels that carry essentials, including

gas pipes, electrical lines and communications cables in many cities, is required to prevent the disruption of power and phone lines becoming a regular occurrence. Some infrastructure, Hill says, may need to be relocated away from low-lying ground entirely. But in the face of sea-level rise and ever more frequent storms, she is less than confident that everywhere will be able to cope. Some districts will be abandoned, at least temporarily, she says. “Cities like Miami, Honolulu, even parts of New York — Queens, Brooklyn, the Rockaways, Staten Island. There are definitely places that are not ready.”

SPONGE CITIES

Conventional city drains can be overwhelmed by a sudden influx of storm water. In cities such as Philadelphia, technologies that allow the streets, pavements and roofs to soak up and store water and release it at a manageable rate are being used to reduce the burden on existing sewers.



PHILADELPHIA WATER

CAPTURE AND CONTROL

An hour's drive from Berkeley, on the opposite side of San Francisco Bay, is Stanford University — one of the participants in ReNUWIt (Re-Inventing the Nation's Urban Water Infrastructure), a research group dedicated to the development of urban water management. Here, adaptation is the name of the game. Environmental engineer and director of the institute Richard Luthy is focused on how cities can control flooding, particularly that caused by storms.

Conventionally, cities have dealt with storm water by trying to get rid of it as fast as possible, says Luthy. “That means building these concrete channels and conduits that take the storm water to the bay or ocean as quick as we can,” he says. If the rainfall is especially strong, however, the sudden volume of water can overwhelm these drainage systems and lead to flash flooding. Slowing down the flow of water through the system and diverting some away entirely can ease this problem — and possibly even turn troublesome storm water into a resource.

Luthy is working with the Californian cities of Sonoma and Los Angeles to improve their resilience to flooding by constructing storm-water reservoirs. Instead of immediately shunting water away down concrete waterways, these large basins will allow storm water to pool and collect.

As well as reducing storm-water-related damage, the reservoirs make it easier to harness the water as a resource. Vegetation planted in and around the reservoir purifies the water because the pollutants become trapped in the soil or the roots, and contaminants are taken up by tissues or even broken down into less harmful substances. The cleaner water can then be further purified and allowed to percolate into underground aquifers beneath the city, where it can serve as a drinking-water source during droughts. “We see storm water as an opportunity, not as a problem,” said Luthy.

A number of cities are including plants and vegetation in water-management plans. Philadelphia's Green City, Clean Water programme, adopted in 2011, uses a variety of techniques to control storm water, including reservoirs like Luthy's. “There's dozens of technologies you can employ,” said Chris Crockett, deputy commissioner of Philadelphia's water department.

One such tool is the tree trench: a deep ditch alongside the pavement, filled with absorbent fabrics, gravel and soil, and then planted with trees (see ‘Sponge cities’). As well as adding a welcome dash of greenery to the urban environment (see page S56), the trench acts as an intermediary step between the storm drain and the sewer, using the same principles of slowing, capturing and purifying runoff to protect the city. The advantage, however, is that the tree trench can be installed in areas where a full-size reservoir could never fit, such as along busy streets.

Roads and streets are conventionally paved with materials such as asphalt or concrete, which are impervious to water and cause it to collect on the surface and flow into storm drains. But engineering can make a difference here too. Philadelphia is experimenting with pervious paving, which allows water to soak through a porous surface to stone reservoirs below, where it can be captured and stored.

Even the roofs of buildings can be used for flood-protection measures. Green roofs use plants and soil to capture and slow storm water, allowing much of it to evaporate before directing the filtered excess into the city's drainage system.

This drive towards newer, cleaner infrastructure to control flooding is not just a US phenomenon. Faced with drainage infrastructure that has not kept pace with rapid urbanization, the Chinese government approved the creation of 16 ‘sponge cities’ in September 2015. These include parts of Chongqing, a city of nearly 10 million people. China hopes to use many of the same technologies being developed in California and Philadelphia, to quickly drain, store and recycle storm water.

Replacing and augmenting outdated sewerage systems and improving a city's resilience to flooding does not come cheap — China will spend about 400 million renminbi ($60 million) on each of its sponge cities per year for the next 3 years, and Philadelphia's programme will need about $2 billion to fund its entire 25-year programme. The price is worth it though, says Crockett. If sea levels rise and the city fails to adapt, it is predicted that annual flood losses could rise from $89 million in 2005 to more than $1 billion by 2050.

Through collaborations such as the C40 Cities Climate Leadership Group, a network of more than 80 cities around the world that together represent more than half a billion people, urban areas have led the way in taking steps to minimize and adapt to climate change. With half the population worldwide in the care of cities, how they adapt to a shifting climate will be crucial to future global health. As seas rise and weather patterns become more unstable, more cities may look to the experiments of Sonoma, Philadelphia and Chongqing for answers. ■

James M. Gaines *is a science writer in Seattle, Washington.*

1. Lin, N., Emanuel, K., Oppenheimer, M. & Vanmarcke, E. *Nature Clim. Change* **2,** 462–467 (2012).
2. Hallegatte, S., Green, C., Nicholls, R. J. & Corfee-Morlot, J. *Nature Clim. Change* **3,** 802–806 (2013).
3. Rotzoll, K. & Fletcher, C. H. *Nature Clim. Change* **3,** 477–481 (2013).

The woodlands of Rock Creek Park in the centre of Washington DC.

JIM LO SCALZO/EPA/CORBIS

GREEN SPACE

A natural high

Exposure to nature makes people happy and could cut mental-health inequalities between the rich and poor.

BY NATASHA GILBERT

Autumn leaves crunch underfoot as I stroll down a path in Rock Creek Park, an urban woodland in Washington DC. I take a deep breath and feel my mood lift. Driving through the busy urban streets to reach the park, I had been worrying about work and what I was going to cook for my family for dinner that evening.

Urban green spaces have value beyond their beauty and environmental importance. Nature improves mental health — people are less depressed when they have better access to green spaces. The beneficial effect is not just a matter of physical exercise, although that is part of the picture. There is something about natural environments that improves people's well-being, says Richard Mitchell, an epidemiologist at Glasgow University, UK. Put simply, being in nature feels good.

Researchers and policymakers are increasingly interested in the link between green spaces and mood because of the implications it could have for preventing and treating mental-health problems in society, says Hannah Cohen-Cline, a researcher at Providence Health and Services in Portland, Oregon. Spending time outdoors in natural environments not only improves people's mental health, but it could also help to reduce health inequalities between the rich and the poor. "Being around nature makes people feel better mentally. This has important policy implications," says Cohen-Cline.

Poor mental health is one of the biggest public-health problems in Western nations. For instance, the Organisation for Economic Co-operation and Development calculates that mental-health conditions such as depression cost the United Kingdom £70 billion (US$100 billion) annually in health-care spending and lost productivity. And the epicentre of these problems is cities: a 2010 meta-analysis showed that urban dwellers are roughly 20% more likely to develop anxiety disorders than their rural counterparts, and nearly 40% more likely to develop mood disorders[1].

DIGGING FOR EVIDENCE

Improving access to green space — such as parks or gardens incorporated into housing developments — in cities could help to cut urban stress, improve city dwellers' mental health and reduce the strain on health-care systems. But until recently, most studies that showed a link between green space and mental health were small, short term and involved groups of similar people, such as students.

"These studies have major limitations," says Mathew White, a social and environmental psychologist at the University of Exeter, UK. It's not clear whether the results are applicable to wider populations or that the beneficial effects persist over time, he says. This is problematic for policymakers who want to see the benefits before investing in health and social interventions.

Scientists are working to tackle these limitations and strengthen the evidence base.

INDIAPICTURES/UIG VIA GETTY IMAGES

White and his colleagues were the first to study changes in mental health over several years as people moved within urban settings. They found that when people moved to areas with more green space, including tree-lined streets, private gardens and public parks, they were happier for at least three years after their move, and that this feeling of contentment grew over time[2].

The research ranked movers' well-being using the short-form General Health Questionnaire (GHQ; a standard clinical tool for measuring anxiety and depression on a scale of 0 to 12). White used an inverse of the GHQ scale so that higher scores represented better mental health. The findings showed that when people moved to areas with more green space, their average GHQ score rose from 9.8 two years before their move to 10.1 three years after their move.

The durability of the happiness effect surprised White. He expected the boost to be short-lived because "people adapt to things quickly." Winning the lottery, for instance, typically makes people happier for up to one year, he says. The benefits of moving to greener areas may last even longer than 3 years (the team only looked at 5 years of data in total) — White is planning a larger study to find out.

White acknowledges that, despite lasting longer than expected, the benefits to mental health seem small. Moving to a greener area is only around one-tenth as important for people's happiness as becoming employed, and has one-third of the impact that marriage does, he says. But, White points out, green space has a greater effect on happiness than low crime rate, which is often cited as a key determinant of well-being. And if the small effect of a green space is multiplied by the thousands of people who use it, that adds up to "a large public-health impact," he says.

Until recently, most studies had been unable to control for the genetic variation that sees some people respond more positively to green space, so it has been difficult to definitively say whether the benefits are due to the green space or to a person's genetic makeup. But Cohen-Cline has unpicked the drivers of mental health using twins. Because twins share at least half of their genes, and those who took part in the study were raised in the same environment, the researchers were able to control both the genetic and environmental factors[3]. "This is important because we know that genetics and childhood environments play a key role in the risk of developing mental-health issues," says Cohen-Cline.

The authors found that green spaces have a direct mental-health impact. People with better access to green space had slightly fewer depressive symptoms than those in less green areas. Independent of any potentially confounding factors, such as childhood environment and genetics, "there is something about green space itself that benefits people's mental health," says Cohen-Cline. Although the twin study shows that green spaces make people happier, it does not say how this works. "It is doing it through several different pathways, and we are still trying to tease that apart," she says. Exercise is known to improve mood, but Cohen-Cline's study found no evidence that it substantially changed people's depressive score, "suggesting that it is not what is driving the association."


Green spaces improve mental and physical health.

One route could be that parks allow people to socialize, which in turn improves their mood. "Social ties are very strongly associated with mental health," says Cohen-Cline. Mitchell is putting his money on another route — people's perception of nature causes physiological changes, such as reducing the stress hormone cortisol and lowering blood pressure. "You perceive nature with your senses," Mitchell says. "Your brain processes those sensory experiences and triggers physiological responses."

Evidence for why this would be is so far thin, but theories abound. One possibility is that people's brains are overexposed to stressful stimuli such as noise and overcrowding in urban environments. By contrast, Mitchell says, natural environments give the brain an opportunity to recover from mental fatigue. It's also possible that our evolutionary heritage means we are simply hard-wired to respond positively to the green spaces that our ancestors grew up in. "We're faced with stressful, noisy environments. When we encounter an environment that is more in keeping with our evolution, that we might innately perceive as more supportive, our bodies and minds react favourably; we literally relax", says Mitchell.

Whatever the underlying explanation, there is evidence that green spaces elicit a direct physiological response, says Mitchell. In Japan, for example, people who spent time participating in Shinrin-yoku — sitting or walking in a forest — had lower cortisol concentrations, pulse rates and blood pressure than when they visited the city[4]. And it's not just parks and forests; blue spaces such as the sea, canals and lakes may give an even bigger boost to people's mood[5].

GREEN INTERVENTION

As the evidence grows, policymakers will be able to design health interventions that use natural resources. The therapeutic and societal value of green spaces is already starting to draw attention. "Policymakers are taking on the message that they have a resource that might be good for people's health and well-being," says Mitchell. The £8.9-million restoration of Clissold Park in northeast London in 2011, for instance, was highlighted by the UK government agency Public Health England in 2014 as an example of a local health intervention.

But creating the spaces isn't enough — says White, there are strong "psychological barriers" that prevent some people from using green space. Just 40% of the UK population will spend time near nature in any given week, and although a lack of time is the main reason given, he says, others say they don't enjoy spending time outside, or that it's not part of their culture.

To engage those most in need, White thinks that health services should offer people with depression 'green prescriptions', which would encourage them to join walking groups or allotments, for example. Physicians could offer this before or as well as drug treatments. White is attempting to work out how these green prescriptions could work in practice, and the potential cost saving for health services.

If participation can be improved, one area that may benefit the most is health inequality. Contentment is not evenly distributed across the socio-economic spectrum: affluence is generally associated with greater happiness. But evidence is beginning to show that green spaces could narrow this gap. Mitchell and his colleagues found that access to green spaces could reduce inequality in mental well-being by 40%[6].

"It is a sizable reduction. Nothing else governments have tried has really had much impact," says Mitchell. The study has its limitations — despite the strong association between green space and decreasing mental-health inequality, there's no proof of causation — but Mitchell is clear: "having a park in your neighbourhood has a greater benefit on poorer people."

"The effects are largest in poorer communities," says White. "Rich people are healthy already." ■

Natasha Gilbert *is a science writer based in Washington DC.*

1. Schoevers, P. *et al. Acta Psych. Scand.* **121,** 84–93 (2010).
2. Alcock, I., White, M., Wheeler, B., Fleming, L. & Depledge, M. *Environ. Sci. Technol.* **48,** 1247–1255 (2014).
3. Cohen-Cline, H. *et al. J. Epidemiol. Community Health* **69,** 523–529 (2015).
4. Park, B. J. *et al. Environ. Health Prev. Med.* **15,** 18–26 (2010).
5. MacKerron, G. & Mourato, S. *Global Environ. Change* **23,** 992–1000 (2013).
6. Mitchell, R. J., Richardson, E. A., Shortt, N. K. & Pearce, J. R. *Am. J. Prev. Med.* **49,** 80–84 (2015).

Nancy Ybarra is one of the caretakers of Pogo Park — a community project in Richmond, California.

STRESS

The privilege of health

Deprivation leads to stress, and stress to bad health. A park, and the science behind it, aims to break that chain.

BY AMY MAXMEN

PRESTON GANNAWAY

Children play on foam mats bathed in late afternoon light as a few staff at Pogo Park hang Christmas lights. Kids squeal as they slide along a zip wire, and a few teenage boys bounce a basketball against concrete painted in pastel designs. It's hard to believe that this is the Iron Triangle — the most deadly part of Richmond, California, one of the most dangerous cities in the United States.

One of the park's caretakers, Nancy Ybarra, now 26, played in the park as a child when it was a dirt patch with two tire swings. Men rolled dice under the trees by day, drank beer and took drugs at night. Ybarra remembers days when police tape surrounded the park after shootings. But the park is safe now. It looks like "a park that could be in a white neighbourhood," Ybarra says. As we walk along the path that encircles it, Ybarra describes how the park is more than just well designed. It was built, and is now staffed, by people from the neighbourhood, and it serves a variety of functions, ranging from employment and childcare to empowerment.

These might seem like social issues, but the park has been justified as a physical-health intervention. In Richmond, as in poor cities around the world, high rates of heart disease, asthma, diabetes and infant mortality correlate with certain social and economic traits. In 2010, unemployment in Richmond was 20%, 38% of the children were living in poverty and the city had the sixth highest crime rate in the United States.

Statistics such as these correlate with negative health outcomes in cities around the world. More than 25% of Richmond's children have asthma, compared with a Californian average of 15%. About 40% of the city's children (and 62% of adults) are overweight or obese, compared with the state's average of 30% for kids. A child born in Richmond is expected to die 11 years earlier than one in the richer city of San Francisco just across the bay. For these reasons, says Jason Corburn, an urban planner and epidemiologist at the University of California, Berkeley, "zip code matters as much as genetic code in understanding a person's susceptibility to disease and premature mortality."

Corburn focuses on how to improve the health of city residents. He argues that cities must transform their physical and social environments — everything from their parks and housing, to the way that community members participate in their city's governance — to improve the health of citizens. Richmond is a short drive from Corburn's office at Berkeley, and the city council has been receptive to his ideas. One of the council's earliest moves was to support local entrepreneur Toody Maher to build Pogo Park. And in 2014, the city enacted a 'Health in All Policies' ordinance. This policy strategy required all municipal decisions to take into account both the physical aspects of health, such as safety and food availability, and social-health issues such as stress.

TOXIC STRESS

For the past 20 years, researchers have worked to solidify the links between chronic psychological stress and physical health. In 1996, Steve Cole, a genomics researcher at the University of California, Los Angeles, noticed that HIV progressed faster in gay men who concealed their homosexuality than in men who were open about it[1]. Cole and his colleagues found that men who concealed their sexuality were

PRESTON GANNAWAY

more psychologically sensitive to social threats, and that stress correlated with increased activity of the sympathetic nervous system (SNS). The basis of the body's fight-or-flight response, the SNS kicks into action by releasing adrenaline and noradrenaline, priming the body for physical activity. Over the next decade, Cole and others described how SNS activation also shuts down proteins involved in fighting viruses, called interferons, allowing HIV to replicate without impediment.

The finding tempted Cole and Greg Miller, a psychologist at Northwestern University in Evanston, Illinois, to explore the connections between stress and other diseases besides HIV. They were particularly interested in subtle, chronic stress, as opposed to the panicky reaction triggered by, say, a near-death accident.

Epidemiologists knew that asthma was disproportionately common in children from poor urban communities, even when family history and air quality were controlled for. To find out whether stress played a part in the asthma prevalence, Miller and his colleagues asked children about their family relationships, friends, school life, home life and neighbourhood. Children with asthma from poorer backgrounds, and in more stressful situations, tended to have higher counts of white blood cells called eosinophils[2]. Given the right allergic trigger, eosinophils kick off the production of molecules that constrict the airways as well as mucus production, resulting in shortness of breath. And eosinophil recruitment, the researchers found, follows SNS activation and the subsequent production of the inflammatory molecule interleukin-5.

Biological connections to chronic stress have also been found in conditions such as cardiovascular disease and type 2 diabetes. People from low socioeconomic communities regularly show signs of chronic stress — levels of stress-related hormones such as cortisol are often higher, for instance. Cole suspects that the chronic stress associated with poverty comes from a lack of certainty about what the future holds. "If you have economic resources, you might develop a blind faith that you can figure out problems," he explains. "But without those resources, there's a baseline of uncertainty."

"We don't have time to wait. This is urgent."

Stress can stem from a lack of control. In post-industrial cities in Finland, researchers have suggested that shame and low self-esteem arising from high rates of unemployment account for the heavy burden of health issues in these areas compared to the country's wealthier cities. For Finnish men who were unemployed twice over a 3-year period, the small risk of death in the 3 years that followed was 168% greater compared with men who had been unemployed only once[3].

Psychological stress among the urban poor is distinct from that of those who live in rural areas, perhaps because income disparity is not as obvious. "Cities expose you to inequality. You see what apartments in rich neighbourhoods look like, which you cannot afford, you know how much food costs at restaurants in those places," says Eldar Shafir, a behavioural scientist at Princeton University in New Jersey. "Well-being is heavily impacted by comparisons," Sharif adds. "It impacts your evaluation of self-worth and self-identity."

Pogo Park could help to improve physical health.

REDUCING THE BURDEN

In poor areas of Richmond, there is no shortage of sources of stress. Violent crime makes a stroll daunting. Poverty means that people worry about where their next meal will come from. Undocumented citizens are on alert because of their precarious status. Institutional racism can lead to feelings of inadequacy and self-defeat. With stressors such as these at play, Richmond's policymakers have realized that medical approaches, such as access to asthma medication, alone will not improve the health of Richmond. Gabino Arredondo, Richmond's Health and Wellness Coordinator says, "when it comes to health, people usually think about doctors and clinics, but we're focusing on upstream interventions."

To understand what stresses residents, Corburn, Arredondo and their team combed through surveys of about 4,000 residents collected since 2007. They also spoke with community-based organizations and analysed geographical data. Areas where violence occurs, car accidents happen and supermarkets are rare were associated with stress. A wide variety of interventions were needed to address the many issues at play. As a result, Richmond has prioritized measures such as speed bumps, lead removal in houses, training to address implicit bias for police officers, and events that help residents to enrol in health insurance.

It may take a decade for Richmond's comprehensive approach to have a demonstrable effect on disease — particularly for conditions such as cardiovascular disease that take years to develop. However, Arredondo argues that if policymakers hold out for this type of evidence before taking action, the biomedical repercussions of letting another generation grow up in stressful conditions will be costly. "We don't have time to wait," Arredondo says. "This is urgent." So the city is plunging ahead.

As it proceeds, the team keeps track of quantitative indicators, such as the rate of asthma in children, as well as survey data that assess residents' opinions over time. The results have provided some glimmers of hope. For example, Corburn says that according to the 2015 surveys, people who identified themselves as belonging to a non-white group and those of a low socioeconomic status reported better perceptions of safety, greater economic and recreational opportunities, and felt more included in the city compared to 2009. He has shared these results with the World Health Organization, and hopes that Richmond's approach will provide insight that policymakers around the world can use to improve the health of the urban poor.

Maher initially pitched her dream of Pogo Park to members of Richmond's community and to employees of the city, such as Arredondo. Collectively, they successfully made the case to the state of California that funding the park could help to ameliorate the isolation, low self-esteem and uncertainty that diminishes the health of Iron Triangle residents. Unlike previous Richmond green-space projects, which had fallen into disarray, Pogo Park was designed and constructed by residents. Today, they watch over it. Eddie Doss, a 59-year old who lives across the street, keeps an eye on the park at night: "If people come here to drink beer, I tell them no, that's against the rules, and they say, OK Eddie."

This year, Richmond secured a US$6.2 million grant from the state to build just over 3 kilometres of safe streets that connect Pogo Park to another park in the Iron Triangle. Pogo Park's success helped Richmond's proposal to beat those from competitors in other cities. Residents who work and volunteer at the park could not be happier with the win — although they, more than anyone, understand that it's just one of many components needed to reduce their burdens. "This park will not solve violence," says Ybarra. But, she adds, "I hope that the kids I take care of here will not become that shooter, they will not become that dope dealer. I'm watching them grow up well." ■

Amy Maxmen *is a freelance science writer based in Berkeley, California.*

1. Cole, S. W. *et al. Psychosom. Med.* **58,** 219–231 (1996).
2. Chen, E. *et al. J. Allergy Clin. Immunol.* **117,** 1014–1020 (2006).
3. Martikainen, P. T. & Valkonen, T. *Lancet* **348,** 909–912 (1996).

PERSPECTIVE

QUEEN'S UNIV. BELFAST

City farming needs monitoring

Pollution poses a significant challenge to food production in urban environments, says **Andrew A. Meharg.**

People are moving from the countryside to cities in ever increasing numbers. As towns and cities grow, farmland is being concreted over. Urban agriculture is often considered to be the future, at least for some components of a city dweller's diet. But cities are far from prime agricultural land, and it cannot be assumed that food produced in urban areas is safe to eat.

Growing food in urban environments seems like an attractive proposition, partly because of the inherent sustainability: waste heat that all cities generate can be harnessed, and grey water (waste water from baths, showers and kitchen appliances) or surface run-off and nutrient-rich sewage effluent can be recycled. As well as a sustainable use of brownfield sites, it can reduce the carbon footprint of food transport and can make cities greener. For poorer families, urban farming produces an income and can diversify diet. Community-based projects can promote social interaction and outdoor activity for a double dose of health benefits. Areas designed with urban farming in mind — such as vertical farms (growing plants up the sides of buildings, for example) and patchworks of fields between, on top of or within blocks of buildings — could shape our future cities. Rooftop and indoor farming would further increase the land area available for agriculture.

But before urban farming can be expanded on a large scale, a key issue that differentiates cities from the countryside must be considered: pollution. Current and past industrial activity, dense and dirty transport infrastructure, domestic burning of fossil fuels and the plethora of chemicals released into domestic waste streams all pollute the soil of cities. Urban sewage and grey water carry detergents and excreted drugs — concerning because many of these are hormonally active — as well as a large suite of industry-derived contaminants. In addition to contaminants from suspended soil dust, city air already has elevated levels of nitrogen oxides, sulfur oxides, hydrocarbons and particulates from car exhausts. Air pollution is known to reduce urban crop yields[1], but the consequences of ingesting foodstuffs covered in these pollutants are not well understood.

Urban soil pollution is not uniform across a city — some areas are severely contaminated, whereas others are cleaner and more suitable for farming. But, in general, fruit and vegetables produced in city environments contain more undesirable substances than rural produce. Whether this increased pollutant burden constitutes a health risk is a matter of debate[2]. However, to ensure that commercially produced food from urban farming is safe, systematic monitoring must be in place. Each potential location should be screened for contaminants and each variety of produce grown must be analysed, because of the differences in the accumulation of contaminants. For example, an extensive survey of fruit and vegetables in a historic mining region of southwest England showed that soil-dust-contaminated leafy vegetables and tuber crops contaminated by direct soil contact were important sources of toxic metals in commercial horticultural produce[3]. Monitoring of this kind is intensive, but the consequences of unsafe produce entering the food chain impel us to implement these measures for crops grown in cities.

Pollution in urban farming flows in both directions. Just as the city can contaminate agricultural produce, farming itself can introduce unwanted chemicals into the environment. Water supplies can be polluted by inorganic fertilizers and manures, which lead to excessive build-up of algae and aquatic plants in nutrient-rich waters, as well as by pesticides. Noxious smells, excessive noise and the stirring up of soil dust into the atmosphere while working the land are all by-products. And having farmland in close proximity to the public is problematic, particularly for infants who may ingest the soil.

One way to allow intensive agricultural activity in urban centres, but to avoid potentially contaminated soil, is hydroponics — growing plants without soil, in water. However, as many cities struggle to supply water for domestic and industrial use, large-scale hydroponic farms would be a further burden on a precious resource. Although in principle hydroponic water could be recycled, this is not always feasible, owing to the water's high nutrient and, if used, pesticide content. In hot climes, bringing water for agriculture into the centre of cities can lead to increases in pests such as malaria-carrying mosquitoes[4].

IN EXISTING CITIES, WHERE URBAN **FARMING** IS AN AFTERTHOUGHT, SOME **LATERAL** THINKING IS REQUIRED.

It may be possible to build new cities that avoid the current contamination issues. But in existing cities, where urban farming is an afterthought, some lateral thinking is required to give urban agriculture a future. Growing non-food crops such as textile fibre plants, biomass crops and timber would make use of urban and suburban waste land, green the city, recycle waste water and biosolids, and produce crops that currently take up rural land that is ideal for food production. Whether farming in cities is cost-effective or not, the non-economic returns such as better living spaces that facilitate social interaction through community-based activity must be considered. If urban waste resources can be recycled into energy, building materials and clothes through farming, everyone will benefit, ultimately making our cities healthier environments. ■

Andrew A. Meharg *is a plant and soil scientist at Queen's University Belfast, UK.*
e-mail: aa.meharg@qub.ac.uk

1. Thomaier, S. *et al. Renew. Agr. Food Syst.* **30,** 43–54 (2015).
2. Leake, J. R., Adam-Bradford, A. & Rigby, J. R. *Environ. Health* **8,** S6 (2009).
3. Norton, G. *et al. Environ. Sci. Technol.* **47,** 6164–6172 (2013).
4. Afrane, Y. A. *et al. Acta Trop.* **89,** 125–134 (2004).

Madina market in Conakry, Guinea. Densely populated urban environments are ideal for the spread of infection.

EMMANUEL BRAUN/REUTERS

DISEASE

Poverty and pathogens

The growth of slums in the developing world's rapidly expanding cities is creating new opportunities for infectious disease to flourish and spread.

BY MICHAEL EISENSTEIN

As Lee Riley read article after article about the deadly 2014 Ebola outbreak, his frustration mounted. "I was seeing all of these newspaper reports and even scientific reports talking about this unprecedented epidemic in West Africa," says Riley, a specialist in urban public health at the University of California, Berkeley, "and there wasn't a single mention of the words 'slums' or 'informal settlements.'"

Ebola is feared because of its high mortality and limited treatment options, but generally it has been limited to remote rural regions. The 2014 outbreak was different: flare-ups in cities such as Conakry in Guinea and Monrovia in Liberia revealed the havoc that this lethal virus could wreak in urban environments. The dense and highly mobile populations provided greater opportunities for the infection to spread. And according to Mosoka Fallah, an epidemiologist who was working with Liberia's Ministry of Health at the front line of the Monrovia outbreak, urban slums bore the brunt. "Wherever there were big outbreaks, most people being infected were among the poor," he says. "Those that didn't have basic sanitation, who had the most distrust of institutions — they also had the most disease."

Developing nations have experienced an astonishing boom in urbanization in the past few decades. The urban population of Kenya, for example, has grown at an average rate of 4.3% per year since 2010, as rural citizens have moved to cities in pursuit of new opportunities. "The range is between 3% and 6% in most of Africa," says Robert Breiman, an infectious-disease epidemiologist at Emory Global Health Institute in Atlanta, Georgia. Riley says that there has been a similar trend in Brazil, where 85% of the population now lives in cities. Many migrants initially make their home in informal settlements at the city periphery. The United Nations Human Settlement Programme UN-Habitat estimates that 863 million people — one-third of the developing world's urbanites — live in slums.

Although better access to medical care means that the health of city dwellers across the socio-economic spectrum is generally superior to that of their rural counterparts, cities can also provide greater opportunities for infectious diseases to flourish. Crowding and poor or non-existent infrastructure exacerbate the risk of infectious disease to slum inhabitants in particular. However, as demonstrated by the rapid spread of the Zika virus over the past year, outbreaks can be a threat to entire cities, nations and — thanks to globalization — the rest of the world.

CLOSE QUARTERS

Breiman has worked extensively in Kibera, one of the oldest and most established slums in Nairobi. Some attempts have been made to assimilate this area into the capital's infrastructure — with mixed results. "There is piped

water that comes in, but it is tapped into by local water sellers," he says. The damage they cause to the pipes makes it easy for contamination to occur. "There's a lot of running sewage, and oftentimes it mixes with the corrupted water pipes." Less established slums have no piped water or waste disposal at all.

These conditions are ideal for the spread of diseases such as cholera, a bacterial diarrhoeal infection transmitted through contaminated food and water. "We still have lots of epidemics of cholera in Africa, and these are typically associated with sanitation," says Amadou Sall, an infectious-disease researcher at the Pasteur Institute in Dakar, Senegal. Typhoid infection occurs through a similar route. Children in Kenya's informal settlements have a one in five chance of contracting typhoid by age ten, says Breiman. In his view, this problem is entirely attributable to poor living standards. "When typhoid went away in New York City and London, and elsewhere, it didn't go away because of vaccines," says Breiman. "It went away because there were physical changes in these environments, and people were living in a different manner."

The rise of urban Ebola was a product of overcrowding and inadequate medical care, as well as close ties between people who had recently arrived in the city and their home villages. The first case diagnosed in Conakry was contracted at a rural family funeral and only turned symptomatic after the person returned to the city. "When that patient got sick, they went to the hospital and initiated several infections among the health-care workers," says Sall. Instead of going to hospitals, many other people in Conakry went to makeshift medical facilities in spaces that lacked the resources to protect carers or patients, such as peoples' homes; this did more to spread the infection than to contain it. The urban poor often travel far and wide to find work, greatly increasing the area that could be affected by the virus and making contact tracing — finding and monitoring people exposed to infected individuals — a daunting undertaking. "In our latest outbreak, we had over 165 contacts that had spread across almost 60% of the city," says Fallah.

A warning of Chagas disease transmission

IMAGEBROKER/REX SHUTTERSTOCK

COUNTRY BUG, CITY BUG

Unlike Ebola or typhoid, many diseases require an intermediary vector — often an insect — to jump from person to person. For these pathogens, becoming a permanent feature of the urban environment is a protracted process because the vector must first establish a foothold there.

An abrupt shift from an agricultural lifestyle to city living can facilitate this process. Chagas disease, a parasitic infection spread by insects called triatomines, has long plagued rural parts of Latin America. But over the past few decades, it has begun to show up in cities as well. The insects accompany rural migrants — and the parasite-infected animals they bring with them — to mid-altitude cities in the Americas, such as Arequipa in Peru. Here, the primary vehicles for the disease are guinea pigs — a popular staple of the Peruvian diet. The poorest immigrants typically congregate in makeshift, illegal settlements known as invasiones. These poorly-constructed shanty towns have the makings of an ideal home for triatomines, which like to dwell in wall cracks. But opportunities to feed are greater in areas that are more densely populated than the relatively spartan invasiones. Because of this, it is the equally poor, but more established areas of the city nearby — those that have made the jump from illegal camps to legitimate urban dwelling — that face the worst infestations. Arequipa's invasiones are merely the entry point for the insects. "Once people get their land titles, there's this infusion of building materials, dogs, guinea pigs and other animals, and more people — that's the tinderbox that allows the bug to take off," says Michael Levy, an epidemiologist at the University of Pennsylvania in Philadelphia, who has worked in Arequipa for more than a decade. Further improvements to living conditions — less crowding, fewer animals and higher-quality homes — seems to be a deterrent to the insects, says Levy.

Whereas triatomines are restricted to certain parts of the Americas, the mosquito *Aedes aegypti* has become entrenched in cities throughout the tropics, largely driven by Asian and African industrialization. This mosquito, which transmits the viruses responsible for dengue and yellow fever, normally breeds in tiny pools of water such as those found in holes in trees. But the rise of cars and plastics has given *A. aegypti* appealing new options. "Tires that wear out would get discarded in the environment, and they make ideal larval habitats," says Duane Gubler, a specialist in tropical medicine at Duke–NUS Medical School in Singapore. The insects also thrive in the water receptacles used by households without ready access to municipal services.

For one major vector-borne disease, urbanization actually provides some protection. The *Anopheles* mosquitoes that transmit malaria are much more particular about where they raise their young than *Aedes*. "They like clean, sunlit pools of water with vegetation around them, which are not typically found in urban areas," says ecologist and epidemiologist Andrew Tatem at the University of Southampton, UK. By combining detailed satellite data on urbanization in Uganda with household surveys for mosquitoes and data on malaria incidence, Tatem and his colleagues have observed a pattern (S. P. Kigozi *et al. Malaria J.* **14,** 374; 2015). "It seems pretty clear that if you move to an urban area, you're at less risk of getting malaria," he says. There are some exceptions: gardens at the city edge can offer breeding grounds, and a species of malaria-transmitting mosquito (*Anopheles stephensi*) found in India can procreate in small water vessels much like *Aedes* can. "But in the concrete jungle, urban centre, you'll generally get zero breeding," says Mark Wilson, an ecologist and epidemiologist at the University of Michigan in Ann Arbor.

NEIGHBOURHOOD WATCH

The 2014 Ebola outbreak was a trial by fire, testing countries' capacity to deal with infectious disease. There was a stark divide between West African countries that promptly brought the epidemic under control and those that did not. Preparedness was a crucial distinguishing element, according to Breiman. Nigeria in particular benefited from a robust infrastructure for polio eradication, which was rapidly converted into a strategy for eliminating Ebola — the country was free of the disease after 3 months and 19 locally transmitted cases.

Such infrastructure is generally most developed in cities. Better community surveillance and prompt delivery of medical care are key advantages for battling the spread of infection. But even a well-designed rapid response can falter in the slums. Conakry established a dedicated Ebola-treatment centre within a month of the first reported case, but Sall and his colleagues encountered numerous hurdles among disenfranchised residents of the city's poorer districts. "People there tend to mistrust the government and the health system," he says. People kept their illness secret because they were afraid of being isolated from their families, and saw no value in reporting to a hospital if no cure was available. "Some people didn't want to go because they were worried about organ smuggling," says Sall. Fallah says that the situation was similar in Monrovia, with families performing secret — and unsafe — burials of their dead.

Grass-roots efforts made the crucial difference. The initial outbreak in Monrovia raged for months, but Fallah and his team regained control by conducting public meetings to discuss fears and concerns. "Many people brought very important suggestions, and this also brought us in touch with influential members of the

ANDRES STAPFF/REUTERS

community," he says. These connections helped track down slum-dwelling contacts who would otherwise have been missed. It also allowed the formulation of public-health practices that were compatible with the community's ethnic customs. "By getting engaged with religious leaders," says Fallah, "we developed methods to make burials safe while keeping them dignified."

VANQUISHING VECTORS

Community-centred strategies have not always been enough. Eliminating *Aedes* mosquitoes, for example, has often demanded aggressive government intervention. Beginning in the 1930s, US epidemiologist Fred Soper waged a successful war on dengue and yellow fever across Latin America — an effort facilitated by the dictatorial governments in power throughout the region. "The programme had a paramilitary type of structure," says Gubler. Soper and his team were given free rein to check every building in a city for mosquito habitats. In similarly heavy-handed fashion, Chile eliminated the Chagas vectors triatomines in its cities under the brutal regime of Augusto Pinochet with a door-to-door campaign overseen by the military police. "Insecticide sprayers would go to each house, and the *carabineros* would be with them," Levy says. "They'd ask if you'd let them in — and everybody said yes!"

Successful as these strategies may have been, the cost to personal liberty makes them impossible to endorse. Thankfully, in Arequipa, Levy and his Peruvian colleagues have made considerable progress through more reasonable means, particularly in poorer quarters where awareness of Chagas disease is high. Although in the richer districts, which have often had the luxury of ignoring the disease, the reception has been icier. Participation in the triatomine elimination campaign is 80–85% in poor districts, but only 64% in richer neighbourhoods. In these communities, says Levy, "the campaign has gone from being a wonderful public-health benefit to an imposition." His group is exploring incentives to boost participation, including community-leader training and raffles with jackpots that reflect the extent of the community's involvement.

"We still have lots of epidemics of cholera in Africa, and these are typically associated with sanitation."

Arequipa's triatomines are now largely under control, but vectors can quickly return if cities lower their guard. This is especially true of *A. aegypti*, because it procreates more rapidly and spreads more aggressively than triatomines. But convincing people to deal with the insects before they become a problem can be tough. As head of the US Centers for Disease Control and Prevention (CDC) dengue-control programme in Puerto Rico in the early 1980s, Gubler led a community-based effort to train residents to evict the mosquitoes from their back gardens.

A water inspection to look for larvae of *Aedes* mosquitoes, which carry the Zika virus.

The results were disappointing. "Nobody controlled the mosquitoes in their yards or houses until there was an epidemic," he recalls. Ultimately, a multipronged attack may be the only real solution — thorough habitat eradication, sensitive diagnostics, access to durable insecticides and, ideally, a vaccine.

A GLOBAL DIAGNOSIS

Although slums provide fertile ground for infectious disease, the problem seldom remains local. Various species of *Aedes* mosquitoes have spread through the trade routes that link the world's cities. China's investment in urban development in Africa, for instance, has brought the Asian tiger mosquito *Aedes albopictus* to the continent. And once the *Aedes* mosquito makes itself at home, a sufficiently virulent strain of dengue can quickly cause a serious outbreak.

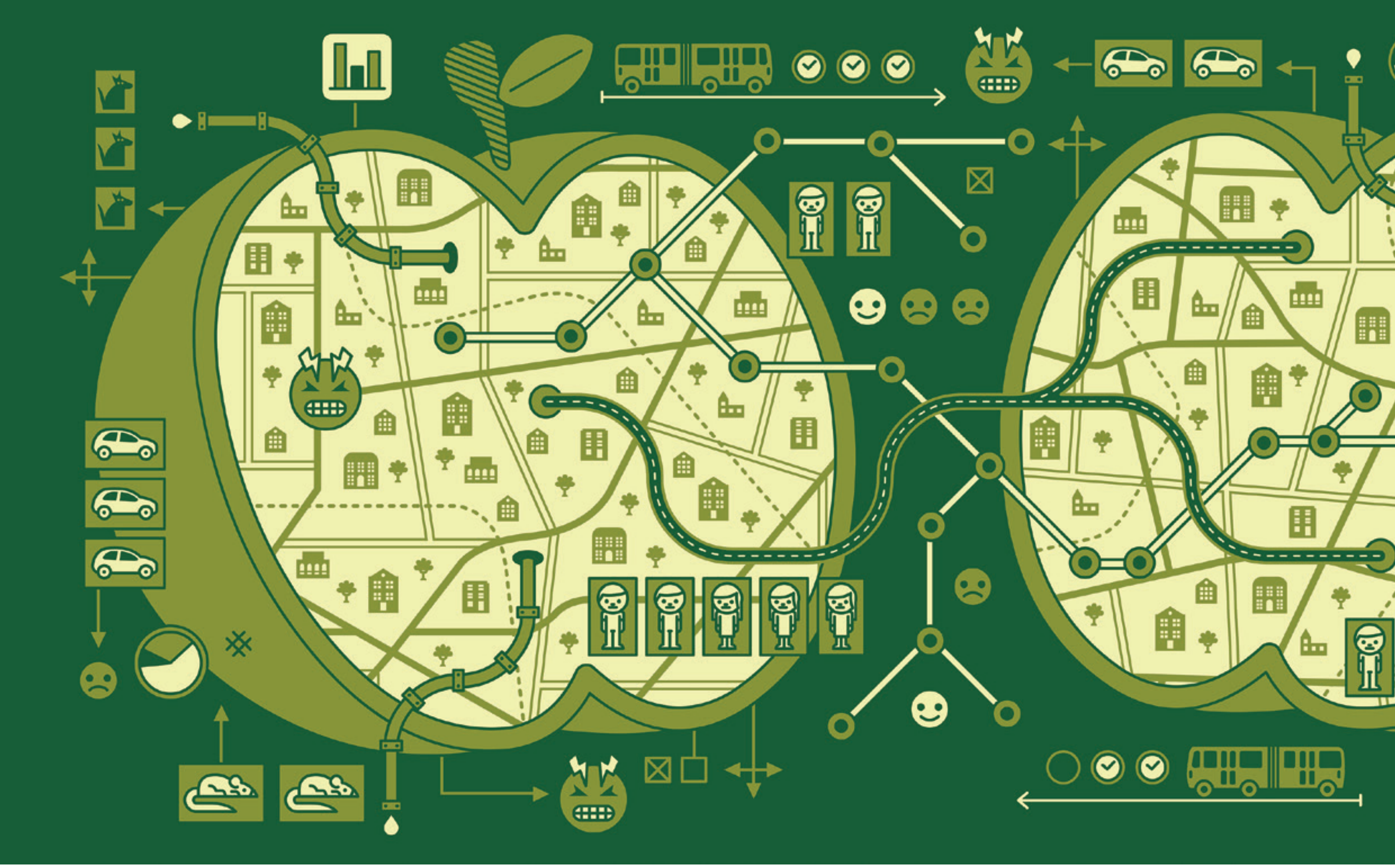
Outbreaks of dengue are facilitated by globalization. Singapore's anti-*Aedes* programmes, which were bolstered after a period of relaxation during the 1980s, have proved inadequate against introduction of the virus by visitors — including large numbers of migrant workers — from endemic regions in south and southeast Asia. A similar process allowed the *Aedes*-borne Zika virus, which was once confined to Asia and Africa, to establish itself in Brazil — one theory suggests that it may have arrived with the influx of tourists for the 2014 football World Cup. The virus has rapidly expanded its reach — the World Health Organization (WHO) is anticipating as many as 4 million cases throughout the Americas. "Zika virus has been around for many decades," says Riley, "but it wasn't until it entered massively urbanized Brazil that it spread like wildfire."

To prevent such propagation, the global public-health community must first intervene to protect the populations at the epicentres of emerging outbreaks. The Ebola epidemic served as an important call to arms. In Liberia, for example, what began as a crisis has turned into an opportunity to develop health-care capacity. "We have built up a competent corps of Liberians who understand the disease well," says Fallah. "We now have ambulances and contact-tracers and lots of partners like CDC, WHO and Unicef working with us." Such efforts could dramatically reduce the severity of future outbreaks.

Nevertheless, the fundamental problem of the slums remains. "If we're going to prevent major epidemics," says Gubler, "we need better living standards — better housing, reliable piped water, good sewage and good hygiene in these tropical urban centres. That's not going to happen any time soon." Some governments have dealt with their slums by simply bulldozing them away. Others have pursued more sympathetic relocation programmes.

But community is essential to quality of life — especially for the urban poor — as Brazil discovered after it transplanted slum-dwellers to high-rises in the 1970s. The government provided a tremendous amount of new infrastructure, but the consequences were not all positive. "This just displaced the social network that these people had," says Riley. This upheaval "engendered crime and other disruption of the social fabric". Over the past decade, says Riley, Brazil's government has focused on improving the slums themselves, providing parks, childcare centres, electricity and drinkable water. In Fallah's view, recognizing this social fabric and partnering with the community leaders who bind it together offers the most cost-effective way to protect urban public health. "You have to have a consistent presence in the community, and not just a sporadic presence whenever there is an outbreak." ■

Michael Eisenstein *is a freelance science writer in Philadelphia, Pennsylvania.*



POLICY

Urban physics

Cities are complex environments. Planning interventions that borrow principles from theoretical physics could help to improve peoples' lives.

BY KEVIN POLLOCK

The ideal city. That was what the French had in mind when they made Hanoi the capital of colonial Indochina at the turn of the twentieth century. To modernize the tropical city, French administrators, technocrats and engineers built railways, bridges and an opera house. But rather than becoming a symbol of success in their southeast Asian colonial project, Hanoi became an example of urban apartheid. The European quarter was renovated with wide, tree-lined avenues, and large, spacious villas. Whereas the old quarter, home to the native Indochinese, was a place of narrow roads and overcrowded buildings.

One part of the modernization plan did not have the intended effect, however. The colonialists had built a system of more than 15 kilometres of underground pipes that flushed waste away in the European quarter — an improvement inspired by the discovery in the late nineteenth century that proper sewers could avert cholera outbreaks by preventing the contamination of drinking water. The old quarter, by contrast, made do with a basic drainage system that dumped untreated waste into the nearby Red River. During the floods, which commonly occurred during the rainy season, the streets in the old city were filled with human faeces.

But the same sewage pipes that spared the colonialists from cholera also served as an ideal breeding ground and mass transportation system for the rat population — harbingers of other serious diseases such as the bubonic plague. Hanoi's wealthiest homes, equipped with indoor plumbing amenities such as running water and flushing toilets, became overrun with the rodents as they left their subterranean homes in search of food. Between 1906 and 1908, plague killed at least 263 people in colonial Hanoi. Bubonic plague remained a risk throughout French colonial rule and the rat problem, although effectively managed, persists today.

By trying to solve one threat to urban health, the French inadvertently introduced another. This pattern has recurred throughout the history of modern cities, as urban policies and interventions perpetually yield unpredicted, and sometimes dire, consequences. This is typical of a network as physically and socially complex as a city. Implementing policies to fix what's broken without introducing a new set of problems is in many ways the central challenge of urban improvement. A way of thinking borrowed from theoretical physics may be the best chance that urban planners have of overcoming it.

COMPLEX MODELS

Cities are complicated. They comprise large numbers of people, and the many ecological, cultural, social and economic entities that make up their environment. All these factors interact in time and space to form complex systems that constantly evolve in response to changes in climate, environment and people.

The characterization of cities as complex systems is based on work that originated in an entirely different field. Physicist and Nobel laureate Philip W. Anderson proposed complexity

JAN KALLWEIT

theory in his 1972 article 'More Is Different' in an effort to understand the "shift from quantitative to qualitative differentiation" in his discipline of many-body physics (P. W. Anderson *Science* **177,** 393–396; 1972). The behaviour of complex aggregates of elementary particles such as the atomic nucleus could not be understood solely in terms of the properties of their individual components. Instead, Anderson reasoned, at each level of complexity entirely new properties appear, and each level of understanding requires a new conceptual structure. An atom in isolation can be thought of as "a featureless, symmetrical little ball", he wrote, but the presence of other particles and environmental conditions changes the shape and reactivity of that atom. Failure to take these changes into account renders the understanding of any measurable observations difficult, if not impossible.

The same is true of cities. As one expands from the perspective of an individual urban dweller to the neighbourhood or town level, new interactions emerge and the complexity of the network increases. Since its inception, Anderson's theory of complex systems has spread beyond particle physics to a variety of other disciplines, including biology and economics.

One of those applying systems thinking to urban environments is Luis Bettencourt, a complex systems specialist at the Santa Fe Institute in New Mexico. Like Anderson, Bettencourt trained as a theoretical physicist — he used empirical data and statistics to model the early Universe. But his fascination with cities and complex social organizations brought him to the institute's Cities, Scaling and Sustainability project, an interdisciplinary effort to develop a general theory of urbanization that is capable of describing cities quantitatively.

"We create models of a city at the systems level," he explains, "in order to describe the physical and social networks of people and businesses, and understand how those mesh with urban space and infrastructure. Just imagine the math of that."

By analysing as much data about cities as possible — everything from crime statistics to efficiency of public transport — Bettencourt and his colleagues have been able to identify key variables and interactions to formulate a picture of an urban system. These models have allowed the team to demonstrate that many of the factors affecting urban well-being are tied, in remarkably predictable ways, to the size of a city; as population grows, so too does crime, for instance. Modelling is also useful in understanding how well-meant health interventions in cities can have unforeseen effects. Although models of complex urban systems can't provide planners with specific policies to enact to improve health and well-being, they can help to explain one of the trickiest properties of complex systems: feedback.

LOOPED THINKING

Complex systems such as cities are alive with feedback loops. The effects of an intervention in one area can produce changes in another, which can, in turn, either amplify or oppose the original intervention. The simplest example of a feedback loop is a thermostat: as the temperature drops below the target level, the heating clicks into life and warms up the room. The increased temperature then feeds back, and the heating is switched off.

In the case of the thermostat, the loop is by design. But in cities, unexpected loops abound. Once again, Hanoi provides a classic example. Faced with the rat problem, in 1902 French colonialists began to pay rat catchers to reduce the population. Wages were dependent on the number of severed tails presented to the authorities. But the rodent bounty did not work as planned — some made the most of the financial opportunity and began breeding rats for their tails.

Feedback loops can also be seen in cities that expand their road networks to tackle traffic congestion. Although the additional road capacity does initially reduce congestion, over time the effect is usually to entice more people to drive, and consequently to bring about more traffic jams, a higher risk of accidents and increased vehicle emissions. Such a loop cannot be described in isolation, however. The effects of the intervention can be seen elsewhere. Where roads are built can affect decisions about housing and energy, and the time spent commuting in traffic can reduce the time spent exercising, increasing the level of obesity (see 'Making connections').

As is the case with any public expenditure, most would agree that urban policy should ideally be evidence based. But with seemingly endless loops of cause and effect to negotiate, determining whether an urban policy has been a success or not can be tricky.

There are many different methods used by the public, private and social sectors of society to evaluate policy impact "which must be adapted to each specific case and each specific problem," says Ernesto Amaral, a sociologist at the RAND Corporation in Santa Monica, California. When measuring the impact of a policy, he says, the ideal scenario is to compare a group that receives the policy with a 'control' group that doesn't. It is a simple enough concept, but as the Hanoi-sewer story illustrates, even in cases in which a control group is available, policies do not have binary outcomes — a single policy can have multiple effects, good and bad.

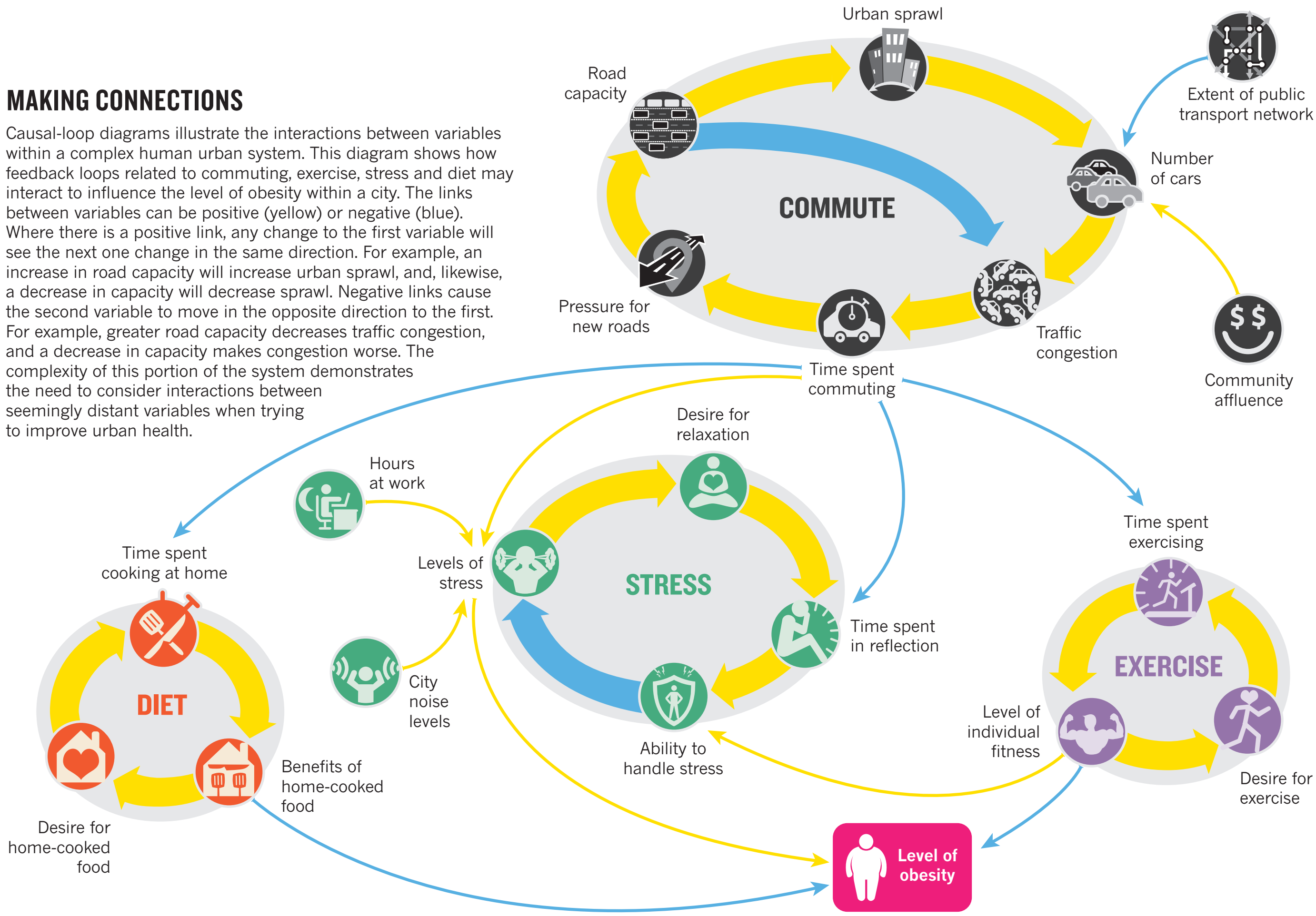
"We're looking at issues of health today, such as those related to dependence on motor vehicles."

Another problem is how to demonstrate that a policy works when the outcomes might take decades or longer to appear. There can be multiple outcomes and the policymakers responsible for the initial intervention may have moved on by the time those outcomes occur. "It gets complicated to correlate back to the specific actions that actually caused them," says Barry Newell, a systems dynamicist at the United Nations University International Institute for Global Health (UNU–IIGH) in Kuala Lumpur, Malaysia.

When faced with having to evaluate such long-term effects, historical studies may prove crucial. "We can use history to understand what has happened as the result of actions in the past, and then project that into the future," says Newell. An example of this type of research is a pilot project run by historian Katrina Proust in George Town, a city on the Malaysian island of Penang. Proust will lead an interdisciplinary team from the UNU–IIGH, including Newell. The researchers will apply the concepts of systems thinking and feedback dynamics to an analysis of almost 200 years of historical archive data — from 1786, when the city was established, to 1985, when the first bridge connecting Penang island with the mainland was completed. Proust hopes to model policy and decision-making behaviour by combining numerical data with well-documented, qualitative observations. The goal is to better understand the relationship between urban administration, planning and public health. "We're looking at issues of health today, such as those related to dependence on motor vehicles," says Proust, "as well as traditional urban problems of the tropics like sanitation and water quality." In general, these types of small-scale, locally focused projects detect feedback effects

MAKING CONNECTIONS

Causal-loop diagrams illustrate the interactions between variables within a complex human urban system. This diagram shows how feedback loops related to commuting, exercise, stress and diet may interact to influence the level of obesity within a city. The links between variables can be positive (yellow) or negative (blue). Where there is a positive link, any change to the first variable will see the next one change in the same direction. For example, an increase in road capacity will increase urban sprawl, and, likewise, a decrease in capacity will decrease sprawl. Negative links cause the second variable to move in the opposite direction to the first. For example, greater road capacity decreases traffic congestion, and a decrease in capacity makes congestion worse. The complexity of this portion of the system demonstrates the need to consider interactions between seemingly distant variables when trying to improve urban health.



SOURCE: BARRY NEWELL

more quickly than large-scale global studies, because response times are relatively short and the people and variables within the system are better identified.

A COMMON LANGUAGE

By identifying the interactions within urban systems, it is hoped that real-world policy can be changed to improve the health of cities. "We want to scientifically understand and model complex systems in cities, to evaluate what the multiple components are and how they interact, in order to predict health and well-being in urban environments," says Franz Gatzweiler, director of the International Council for Science's Urban Health and Well-Being programme based at the Institute of Urban Environment in Xiamen, China. The programme supports research projects that use a systems approach to generate useful products for policymakers, because although projects such as Proust's can improve the understanding of systems, there are few examples of this type of approach being applied.

One group committed to this approach with some success is InWithForward, an organization based in Vancouver, Canada, that combines systems analysis and modelling with engagement of communities at the local level. They identify the shortcomings of existing urban policies by observing the system from the point of view of local individuals, and then work with these people to fill the gaps within the system. For example, by observing people with cognitive disabilities in Vancouver, the group discovered that although the medical care was adequate, most of these people had mundane and repetitive daily routines, and that their quality of life was poorer because of it. The organization's solution was to connect these individuals with other members of their community who could provide diverse opportunities and learning experiences — from art and music teachers, to carpenters and animal trainers. "Most of the data policymakers have available tells them where the problem is, but don't help to illuminate what the solutions would be," says Sarah Schulman, founder of the organization. Rather than major top-down changes, Schulman's approach focuses on creating new interactions and networks within the existing urban system, promoting a shared understanding among both policymakers and urban dwellers of how the system works, and what can be done to change things.

"Our capacity to act collectively is lagging."

But shared understanding is not a simple thing to achieve, especially between urban planners who tend to treat areas of the city separately. "We all talk different languages," says Gatzweiler. Different disciplines have their own vocabularies, and it is entirely possible for miscommunication to pass unnoticed. For example, despite physicist Newell and historian Proust working and living together (the two researchers are married), they only realized that they were each defining a term often used in their work entirely differently after five years on a particular project.

"Our capacity to act collectively is lagging," says Gatzweiler. Overcoming the silos into which different aspects of the city network are divided — water, transport, health and so on — is a central challenge of urban planning. In a system as complex as a city, no one person or speciality can see the whole picture; urban health is an interdisciplinary problem that requires collaboration across administrative and scientific barriers. Once again, Anderson's theory of complexity makes the way ahead clear. Rather than have each discipline focus on making inroads into problems by themselves, Anderson wrote, "we should recognize that such roads, while often the quickest shortcut to another part of our own science, are not visible from the viewpoint of one science alone."

Threats to urban health and well-being span disciplinary boundaries, and systems thinking offers a tool by which these boundaries can be bridged, says Gatzweiler "We need systems thinking and the co-production of knowledge in order to create healthier, wealthier cities, and people within them." ■

Kevin Pollock *is a freelance journalist based in Tel Aviv-Yafo, Israel.*

nature climate change

LETTERS

PUBLISHED ONLINE: 18 AUGUST 2013 | DOI: 10.1038/NCLIMATE1979

Future flood losses in major coastal cities

Stephane Hallegatte[1,2]*, Colin Green[3], Robert J. Nicholls[4] and Jan Corfee-Morlot[5]

Flood exposure is increasing in coastal cities[1,2] owing to growing populations and assets, the changing climate[3], and subsidence[4–6]. Here we provide a quantification of present and future flood losses in the 136 largest coastal cities. Using a new database of urban protection and different assumptions on adaptation, we account for existing and future flood defences. Average global flood losses in 2005 are estimated to be approximately US$6 billion per year, increasing to US$52 billion by 2050 with projected socio-economic change alone. With climate change and subsidence, present protection will need to be upgraded to avoid unacceptable losses of US$1 trillion or more per year. Even if adaptation investments maintain constant flood probability, subsidence and sea-level rise will increase global flood losses to US$60–63 billion per year in 2050. To maintain present flood risk, adaptation will need to reduce flood probabilities below present values. In this case, the magnitude of losses when floods do occur would increase, often by more than 50%, making it critical to also prepare for larger disasters than we experience today. The analysis identifies the cities that seem most vulnerable to these trends, that is, where the largest increase in losses can be expected.

A first screening study[1] provided a global overview of flood exposure in world coastal cities. The exposure metric can be viewed as a worst case scenario, but it does not estimate average annual losses, which is a standard metric in disaster risk management planning. To do so, it is necessary to take into account infrastructure-based adaptation (for example, dykes) and the vulnerability of populations and assets. Here, we assess economic average annual losses (AAL) in 136 coastal port cities, using a method developed for assessing city-level flood risk[7] and a new database of urban coastal protection (Methods).

Present aggregated average annual flood losses in the 136 cities are estimated at approximately US$6 billion per year. Table 1 ranks the most vulnerable cities in 2005 using two different metrics of vulnerability. In the left column, the table shows a ranking in terms of AAL, taking into account all potential floods and existing protection. The AAL estimates can be compared to more sophisticated approaches. For instance, the annual losses in New Orleans are estimated at US$600 million, close to the US$650 million estimates from the Interagency Performance Evaluation Taskforce[8]. In the right column, cities are ranked according to relative vulnerability, namely the ratio of AAL to the city's gross domestic product (GDP). This value can be understood as the share of the city's economic output that should be saved annually to pay for future flood losses. The 20 cities most vulnerable according to this last indicator are also presented in Fig. 1.

The ranking in terms of exposure includes mainly rich-country cities (Supplementary Table S4). On average, however, rich cities are better protected than poorer ones, and the ranking in terms of absolute flood losses contains more cities from developing countries. In relative terms, developing-country cities are even more vulnerable, with only three cities from developed countries in the top 20 (New Orleans, Miami and Tampa—Saint-Petersburg). Moreover the ranking in absolute terms (left column) includes mainly capital cities, whereas secondary cities are more often represented in the ranking in relative terms (right column). This difference suggests that risk management efforts may be lower in secondary cities.

Table 1 shows the importance of existing flood defences: in a city such as Amsterdam, exposure is extremely high (US$83 billion of assets exposed to the 100-year flood), but AAL do not exceed US$3 million, because estimated defence standards are the highest that exist globally. On the other hand, a city such as Ho Chi Minh, in Vietnam, has a 100 year exposure of only US$18 billion, but the lower level of protection means that the city is affected by small floods on a frequent basis, resulting in large estimated average costs. In relative terms, Ho Chi Minh City has one of the largest vulnerabilities, with AAL reaching 0.74% of local GDP. The ratio of AAL to local GDP exceeds 1% for two cities, Guangzhou and New Orleans. The vulnerability of New Orleans has been reduced however by recent post-Hurricane Katrina investments and is likely to be reduced further in the near future[9].

Another conclusion from Table 1 is the concentration of losses in only a few cities. Only 13 cities have average losses in excess of US$100 million, and three American cities (Miami, New York City and New Orleans) explain 31% of the global aggregate losses in the 136 cities, because of their high wealth and low protection level. Adding Guangzhou, the four top cities explain 43% of global losses. Also, the US seems particularly vulnerable, with 6 American cities in exposure ranking, 8 in the ranking by absolute AAL, and 3 in the ranking by relative AAL. As coastal flood risks are highly concentrated, flood reduction actions in a few locations could be very cost-effective.

To develop possible future patterns of drivers of risk to 2070, our analysis introduces three scenarios for socio-economic changes and six for environmental change. From there, we retain four main scenarios: SEC assumes only socio-economic changes, derived from OECD and UN scenarios; SEC-S adds subsidence to scenario SEC (40 cm in 2050 in the cities subjected to subsidence); and SLR-1 and SLR-2 add optimistic and pessimistic sea-level rise scenarios to SEC-S, respectively (with 20 cm and 40 cm in 2050). Here, we report results for 2050, but results for 2030 and 2070 are available in the Supplementary Information.

With no adaptation, the projected increase in average losses by 2050 is huge, with aggregate losses increasing to more than US$1 trillion per year in scenarios SLR-1 and SLR-2 (Supplementary Table S6). All cities experience a similar increase

[1]The World Bank, Sustainable Development Network, Washington DC 20433, USA, [2]Centre International de Recherche sur l'Environnement et le Développement (CIRED), Nogent-sur-Marne 94736, France, [3]Flood Hazard Research Centre, Middlesex University, London NW4 4BT, UK, [4]Faculty of Engineering and the Environment, University of Southampton, Southampton SO17 1BJ, UK, [5]Organisation for Economic Co-operation and Development, Paris 75775, France. *e-mail: shallegatte@worldbank.org

NATURE CLIMATE CHANGE DOI: 10.1038/NCLIMATE1979

Table 1 | City ranking by risk (AAL) and relative risk (AAL in percentage of GDP) for 2005.

Ranking by AAL (US$ million)					Ranking by relative AAL (percentage of city GDP)				
	Urban agglomeration	100 year exposure	AAL, with protection (US$ million)	AAL, with protection (percentage of GDP)		Urban agglomeration	100 year exposure	AAL, with protection (US$ million)	AAL, with protection (percentage of GDP)
1	Guangzhou	38,508	687	1.32%	1	Guangzhou	38,508	687	1.32%
2	Miami	366,421	672	0.30%	2	New Orleans	143,963	507	1.21%
3	New York—Newark	236,530	628	0.08%	3	Guayaquil	3,687	98	0.95%
4	New Orleans	143,963	507	1.21%	4	Ho Chi Minh City	18,708	104	0.74%
5	Mumbai	23,188	284	0.47%	5	Abidjan	1,786	38	0.72%
6	Nagoya	77,988	260	0.26%	6	Zhanjiang	2,780	46	0.50%
7	Tampa—St. Petersburg	49,593	244	0.26%	7	Mumbai	23,188	284	0.47%
8	Boston	55,445	237	0.13%	8	Khulna	2,073	13	0.43%
9	Shenzen	11,338	169	0.38%	9	Palembang	1,161	27	0.39%
10	Osaka—Kobe	149,935	120	0.03%	10	Shenzen	11,338	169	0.38%
11	Vancouver	33,456	107	0.14%	11	Hai Phòng	6,348	19	0.37%
12	Tianjin	11,408	104	0.24%	12	N'ampo	507	6	0.31%
13	Ho Chi Minh City	18,708	104	0.74%	13	Miami	366,421	672	0.30%
14	Kolkata	14,769	99	0.21%	14	Kochi	855	14	0.29%
15	Guayaquil	3,687	98	0.95%	15	Tampa—St. Petersburg	49,593	244	0.26%
16	Philadelphia	22,132	89	0.04%	16	Nagoya	77,988	260	0.26%
17	Virginia Beach	61,507	89	0.15%	17	Surat	3,288	30	0.25%
18	Fukuoka—Kitakyushu	39,096	82	0.09%	18	Tianjin	11,408	104	0.24%
19	Baltimore	14,042	76	0.08%	19	Grande_Vitória	6,738	32	0.23%
20	Jakarta	4,256	73	0.14%	20	Xiamen	4,486	33	0.22%

A comparison with a ranking by exposure is proposed in the Supplementary Information.

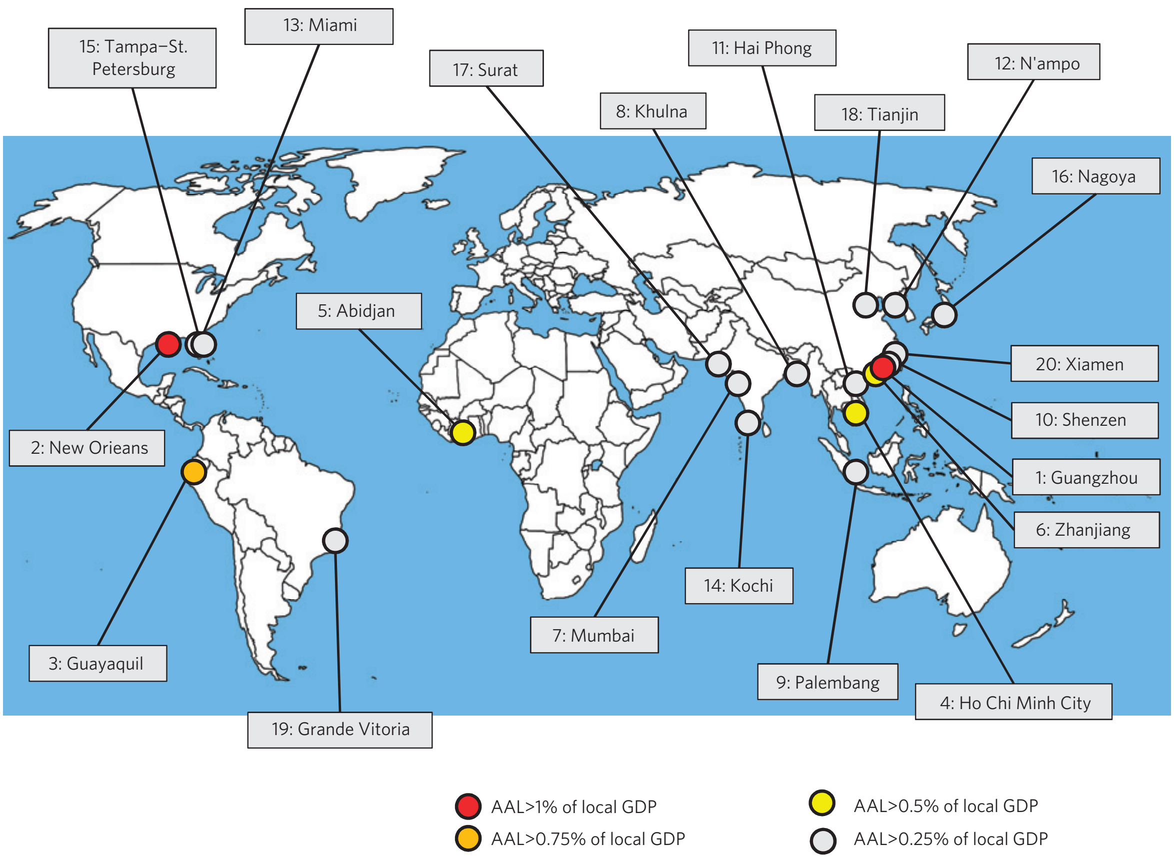


Figure 1 | The 20 cities where the relative risk is larger in 2005, that is, where the ratio of AAL with respect to local GDP is the largest. More information in Table 1.

Table 2 | The 20 cities with the highest loss in 2050, assuming scenario SLR-1 and adaptation option that maintains flood probability (option PD).

Urban agglomeration	Scenarios with socio-economic change alone (SEC)		Scenarios with socio-economic change, subsidence, sea-level rise and adaptation to maintain flood probability (scenarios SLR-1, and adaptation option PD)		
	AAL (US$ million)	AAL (percentage of city GDP)	AAL (US$ million)	Increase in AAL compared with 2005 (%)	AAL (percentage of city GDP)
Guangzhou (S)	11,928	1.32%	13,200	11%	1.46%
Mumbai	6,109	0.47%	6,414	5%	0.49%
Kolkata (S)	2,704	0.21%	3,350	24%	0.26%
Guayaquil (S)	2,813	0.95%	3,189	13%	1.08%
Shenzen	2,929	0.38%	3,136	7%	0.40%
Miami	2,099	0.30%	2,549	21%	0.36%
Tianjin (S)	1,810	0.24%	2,276	26%	0.30%
New York—Newark	1,960	0.08%	2,056	5%	0.08%
Ho Chi Minh City (S)	1,743	0.74%	1,953	12%	0.83%
New Orleans (S)	1,583	1.21%	1,864	18%	1.42%
Jakarta (S)	1,139	0.14%	1,750	54%	0.22%
Abidjan	826	0.72%	1,023	24%	0.89%
Chennai (Madras)	825	0.12%	939	14%	0.14%
Surat	905	0.25%	928	3%	0.26%
Zhanjiang (S)	806	0.50%	891	11%	0.55%
Tampa—St. Petersburg	763	0.26%	859	13%	0.29%
Boston	741	0.13%	793	7%	0.14%
Bangkok (S)	596	0.07%	734	23%	0.09%
Xiamen (S)	572	0.22%	729	27%	0.29%
Nagoya (S)	564	0.26%	644	14%	0.30%

'S' indicates that the city is prone to significant subsidence. Most of these cities are located in deltaic regions, where subsidence influences local sea level in 2050.

in risk. In the absence of adaptation, the impact of environmental change is much larger than the effect of socioeconomic change. These numbers should not be considered as predictions, but they demonstrate the need for adaptation, because inaction would result in unacceptably high losses.

We then consider adaptation and how it will alter losses. We assume first that adaptation action increases coastal flood defences to maintain a constant probability of flooding (adaptation option: present design, PD). The increase in aggregate AAL is much lower in this case. Owing to socio-economic change, there is still a ninefold increase in aggregate losses, from US$6 to US$52 billion per year, but this is made more manageable by the fact that these cities are also much richer. However, rising water levels still increase AAL: subsidence by 12% and sea-level rise by an additional 2–8%, reaching between US$60 and 63 billion per year.

Table 2 shows the top 20 cities in terms of AAL in 2050 in the scenario with subsidence and optimistic sea-level rise (SLR-1), with adaptation to maintain present flood probability. Guangzhou remains the most vulnerable city, with AAL exceeding US$13 billion. With socio-economic change alone, AAL in Guangzhou would be around US$12 billion per year in 2050 (a 17-fold increase in absolute terms). Subsidence and sea-level rise are thus responsible for an additional 10% increase, that is, a 10% increase in the AAL-to-GDP ratio. Indeed, even if the probability of coastal flooding is unchanged thanks to upgraded coastal defence infrastructure, the fact that a larger share of existing assets is protected by these defences means that annual losses will rise relative to local GDP. For instance, sea-level rise and subsidence increase the AAL-to-GDP ratio by 54% in Jakarta and by 24% in Abidjan even if present flood probabilities are maintained thanks to better defences. In other words, the world sees no more floods, but each flood is more destructive owing to sea-level rise and subsidence, even with better defences. This effect reinforces a trend that can be expected from socio-economic change alone, even in the absence of environmental change.[10]

Figure 2 shows the 20 cities where the increase in average annual losses between 2005 and 2050 is greatest in relative terms; detailed numerical values are provided in Supplementary Table S7. In Alexandria, for instance, maintaining flood probability leads to an increase by 154% in AAL. These most vulnerable cities are distributed all over the world, with a concentration in the Mediterranean Basin, the Gulf of Mexico and East Asia. Even though absolute levels of risk are sometimes low in these cities, they can be considered as adaptation hotspots because this is where flood risks are likely to increase the most in relative terms.

To avoid any increase in risk, an adaptation policy needs to do more than maintain present flood probability. Rather, maintaining present levels of risk (relative to local GDP) in the context of rising sea levels, subsidence and socio-economic changes requires adaptation policy that reduces flood probability over time. In the adaptation option termed present losses, an upgrade in defence is thus calibrated to cancel the impact of environmental changes and to maintain present losses on average relative to local wealth, keeping aggregate losses at US$52 billion.

For each city, we estimate the increase in defence standard that would maintain the relative risk level (that is, keep constant the ratio of average annual losses to local GDP). The required increase in protection is larger than local sea-level rise. For instance, in Alexandria, protection needs to be raised by 67 cm, for a 60 cm rise in local sea level; this corresponds to moving from a 100-year design standard to a 270-year design standard (that is, a division by 2.7 of the probability of flooding). In other cities (Supplementary Table S7) the increase in dyke height is between 2 and 8 cm larger than sea-level rise. This increase corresponds to a significant increase in the standard of protection, that is, to a large decrease

NATURE CLIMATE CHANGE DOI: 10.1038/NCLIMATE1979

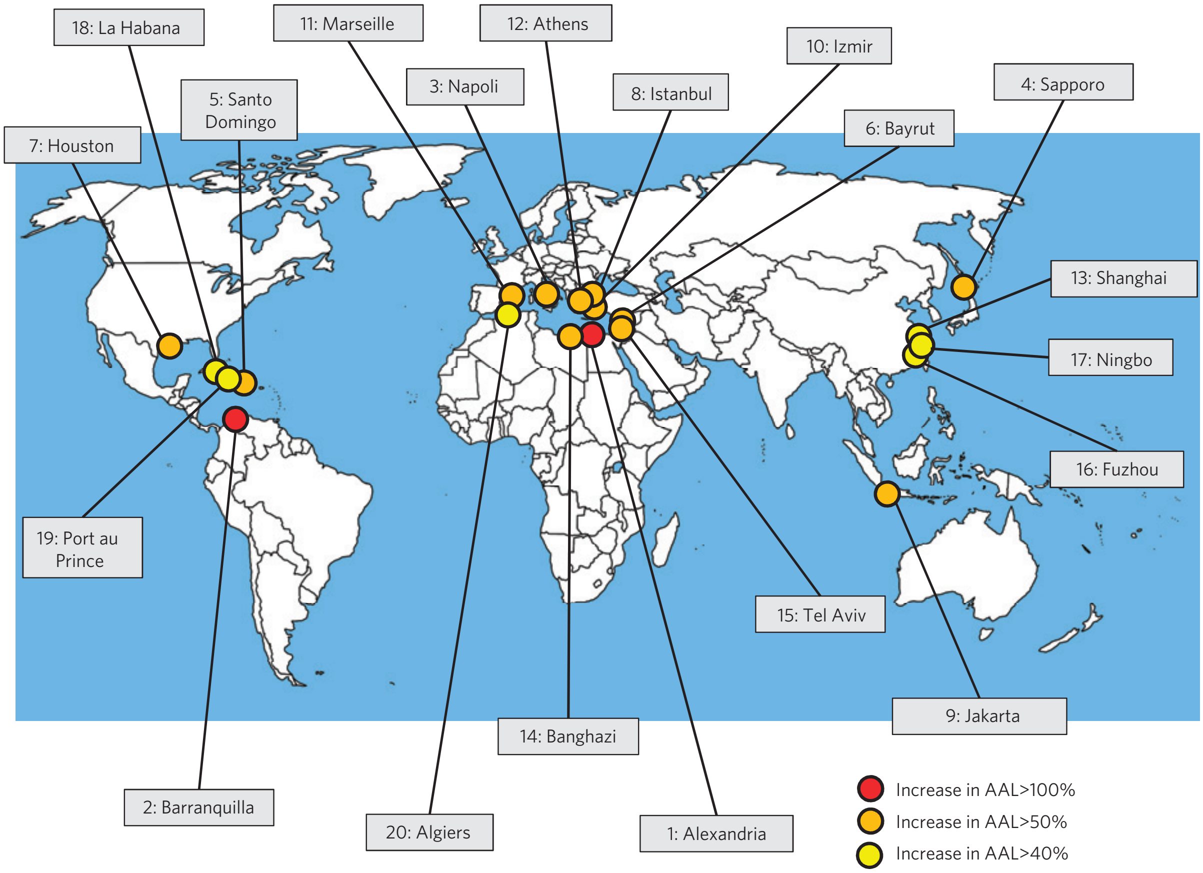


Figure 2 | The 20 cities where AAL increase most (in relative terms in 2050 compared with 2005) in the case of optimistic sea-level rise, if adaptation only maintains present defence standards or flood probability (PD). More information in Supplementary Table S7.

in the probability of flooding (for example, a division by 1.6 in Jakarta and Shanghai).

Even if the relative risk level is maintained, flood consequences would be much higher in a world with sea-level rise and better protection. In Alexandria in 2050, for instance, the probability of a flood may decrease by 2.7 owing to better defences, but if a flood occurs the total losses triple from US$17 billion (with socio-economic change alone) to US$51 billion, a tripling due to environmental changes alone. Results for other cities are presented in Supplementary Table S7. Many cities would experience losses that are more than 50% larger if an event exceeds the post-adaptation protection level. Hence, the world's coastal cities become more dependent on flood defences, but also more vulnerable once failure or overtopping occurs.

Finally, Table 3 looks at present city characteristics that influence the vulnerability in 2050. It shows that cities that grow rapidly, have large populations, are poor, exposed to tropical storms, and prone to subsidence are over-represented in the top 20 for absolute AAL (the cities from Table 2). However, these drivers are not as relevant for the relative increase in annual average loss with adaptation option PD (the cities from Fig. 2). Only subsidence seems to be a consistently good determinant of vulnerability for both absolute and relative measures of change, with twice as many cities with subsidence in the top 20 based on the two indicators (absolute and relative losses). The cities most vulnerable in relative terms may thus not be the ones suggested by the present situation and historic floods, nor are they the ones that necessarily attract the most research and analysis today with respect to managing risk.

While recognizing the limitations and uncertainties in this analysis, three important policy conclusions can be drawn that are robust across a range of plausible assumptions. First, failing to adapt is not a viable option in coastal cities. It is difficult to estimate the cost of these adaptation options, as it depends on the specific context for each city and on selected approaches and technologies. On the basis of anecdotal evidence from a few cities[11–14], a few billion US dollars per city in initial investment—plus approximately 2% of the initial investment cost in annual operation and maintenance costs—is the possible order of magnitude for adaptation costs[15]. Hence, indicative annualized values with 5% interest rates are about US$350 million per year per city, or approximately US$50 billion per year for our 136-city sample. These estimated aggregate adaptation costs are far below our estimate of aggregate damage losses per year in the absence of adaptation, and of the same order of magnitude as residual losses with adaptation (Supplementary Table S6). These estimates include only flood risks and do not encompass all weather risks that these cities face; see, for instance, ref. 16 for a global analysis of wind damage from tropical cyclones.

Second, managing coastal flood risk requires doing more than maintaining today's standard of protection (and present probability of flooding). In practice, probability of flooding will need to be reduced to maintain flood risks at today's levels.

Last, improving standards of protection could maintain or reduce risk levels and decrease the number of floods, but the magnitude of losses when floods do occur will still increase. This result points to the limitations of what infrastructure-based adaptation can achieve. As illustrated by the recent landfall of hurricane Sandy on the east coast of the United States, there is a need to prepare at the local, national and international level for larger floods and the disasters that ensue. Such preparations can include strengthening disaster planning measures, including early warning and evacuation systems, more comprehensive insurance

Table 3 | The fraction of cities in the five categories (fast growing, large population, low income, subject to tropical storms and subject to subsidence) in the full sample of 136 cities, and in two top 20 categories using different indicators.

	Fast-growing city (local GDP growth in 2005 larger than 5% per year)	Large population (larger than 5 million in 2005)	Low income (GDP per capita in 2005 less than US$5,000)	Subject to hurricane and tropical storms	Prone to significant subsidence
All 136 cities	40%	19%	27%	51%	27%
Top 20 in terms of AAL in 2050 (adaptation option PD; Table 2)	65%	55%	40%	95%	55%
Top 20 in terms of increase in AAL (adaptation option PD; Fig. 2)	40%	15%	20%	45%	40%

schemes and other forms of post-disaster response to quickly rebuild affected communities.

Methods

Flood risks are analysed following ref. 7. The population exposure is taken from a previous analysis[1]. Exposed population was translated into exposed assets using an estimate of produced capital per inhabitant drawing on recent work from the World Bank[17]. The DIVA database provides information about extreme water levels[18]. We create a first database for coastal defences; this is based on collected evidence on existing defences where possible, and the authors' expert estimates to complete the defence database (Supplementary Information).

Then, we calculate the probability of different flood levels in each city (within the flood defences) using three simple models for defence failure or overtopping. Even though absolute risk levels depend on the failure model, the relative effect of sea-level rise and subsidence on losses is relatively robust to this uncertainty.

From the probability of the various flood levels in the city, and from data on assets that are exposed at different flood levels, flood asset losses are estimated using depth–damage functions for 6 categories of assets.

To assess future losses, we use two socio-economic scenarios[19]: an OECD-based growth scenario in which urban populations grow at the same rate in all cities, following an extrapolation of UN urbanization scenarios; and the OECD-based growth scenario in which city population is capped at 35 million inhabitants. We assume that future assets in the city have the same elevation distribution as existing assets.

The six scenarios of environmental change combine: two assumptions on subsidence (no subsidence, or a 40 cm subsidence in 2050 in all cities subject to it); and three assumptions about sea-level rise (none, 20 cm, or 40 cm in 2050). Combined with socio-economic scenarios and adaptation options, these lead to 108 scenario combinations. Results for the 108 scenarios and input data and model codes are available in the Supplementary Information. The analysis presented here considers only four scenarios of the total number of scenarios (see text). All scenarios use the simplest defence failure model, the maximum protection level, and constrain cities to no more than 35 million inhabitants.

Received 4 December 2012; accepted 16 July 2013; published online 18 August 2013

References

1. Hanson, S. *et al.* A global ranking of port cities with high exposure to climate extremes. *Climatic Change* **104,** 89–111 (2011).
2. De Sherbinin, A., Schiller, A. & Pulsipher, A. The vulnerability of global cities to climate hazards. *Environ. Urban.* **19,** 39–64 (2007).
3. Nicholls, R. J. *et al.* in *IPCC Climate Change 2007: Impacts, Adaptation and Vulnerability* (eds Parry, M. L., Canziani, O. F., Palutikof, J. P., van der Linden, P. J. & Hanson, C. E.) 315–356 (Cambridge Univ. Press, 2007).
4. Nicholls, R. J. Coastal megacities and climate change. *GeoJournal* **37,** 369–379 (1995).
5. Dixon, T. H. *et al.* Space geodesy: Subsidence and flooding in New Orleans. *Nature* **441,** 587–588 (2006).
6. *Climate Risks and Adaptation in Asian Coastal Megacities* (The World Bank, 2010).
7. Hallegatte, S. *et al.* Assessing climate change impacts, sea level rise and storm surge risk in port cities: A case study on Copenhagen. *Climatic Change* **104,** 113–137 (2011).
8. http://nolarisk.usace.army.mil/.
9. Link, L. E. The anatomy of a disaster, an overview of Hurricane Katrina and New Orleans. *Ocean Eng.* **37,** 4–12 (2010).
10. Hallegatte, S. *An Exploration of the Link between Development, Economic Growth, and Natural Risk* Policy Research Working Paper No. 6216 (The World Bank, 2012).
11. Evans, E. *et al. Proc. Inst. Civil Eng.* **159,** 53–61 (2006).
12. Kabat, P. *et al.* Dutch coasts in transition. *Nature Geosci.* **2,** 450–452 (2009).
13. Kates, R. W., Colten, C. E., Laska, S. & Leatherman, S. P. Reconstruction of New Orleans after Hurricane Katrina: A research perspective. *Proc. Natl Acad. Sci. USA* **103,** 14653–14660 (2006).
14. Ammerman, A. J. & McClennen, C. E. Saving Venice. *Science* **289,** 1301–1302 (2000).
15. Nicholls, R., Brown, S., Hanson, S. & Hinkel, J. *Economics of Coastal Zone Adaptation to Climate Change* (The World Bank, 2010).
16. Peduzzi, P. *et al.* Global trends in tropical cyclone risk. *Nature Clim. Change* **2,** 289–294 (2012).
17. *The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium* (The World Bank, 2010).
18. Vafeidis, A. T. *et al.* A new global coastal database for impact and vulnerability analysis to sea-level rise. *J. Coast. Res.* 917–924 (2008).
19. Chateau, J., Rebolledo, C. & Dellink, R. *An Economic Projection to 2050: The OECD 'ENV-Linkages' Model Baseline* No. 41 (OECD, 2011).

Acknowledgements

These are early results from an ongoing OECD research project. The authors acknowledge support from the OECD for this research and support from J. Chateau to provide the socio-economic scenario data from the OECD's ENV-Linkages model. C.G. acknowledges support from the AVOID project. The views shared in this article represent those of the authors and are not intended to represent the views of the OECD or of its Member countries, or the view of World Bank, its executive directors, or the countries they represent.

Author contributions

The four authors designed the study, interpreted results and authored the paper. S.H. developed and ran the models. R.N. and C.G. provided expert input on depth–damage curves and coastal protection.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.H.

Competing financial interests

The authors declare no competing financial interests.

ARTICLE

Received 12 Oct 2013 | Accepted 28 Mar 2014 | Published 13 May 2014

DOI: 10.1038/ncomms4749

Two-stroke scooters are a dominant source of air pollution in many cities

S.M. Platt[1], I. El Haddad[1], S.M. Pieber[1], R.-J. Huang[1], A.A. Zardini[2], M. Clairotte[2,†], R. Suarez-Bertoa[2], P. Barmet[1], L. Pfaffenberger[1], R. Wolf[1], J.G. Slowik[1], S.J. Fuller[3], M. Kalberer[3], R. Chirico[1,†], J. Dommen[1], C. Astorga[2], R. Zimmermann[4,5], N. Marchand[6], S. Hellebust[6], B. Temime-Roussel[6], U. Baltensperger[1] & A.S.H. Prévôt[1]

Fossil fuel-powered vehicles emit significant particulate matter, for example, black carbon and primary organic aerosol, and produce secondary organic aerosol. Here we quantify secondary organic aerosol production from two-stroke scooters. Cars and trucks, particularly diesel vehicles, are thought to be the main vehicular pollution sources. This needs re-thinking, as we show that elevated particulate matter levels can be a consequence of 'asymmetric pollution' from two-stroke scooters, vehicles that constitute a small fraction of the fleet, but can dominate urban vehicular pollution through organic aerosol and aromatic emission factors up to thousands of times higher than from other vehicle classes. Further, we demonstrate that oxidation processes producing secondary organic aerosol from vehicle exhaust also form potentially toxic 'reactive oxygen species'.

[1] Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, CH-5232 Villigen, Switzerland. [2] European Commission Joint Research Centre, Institute for Energy and Transport, 21027 Ispra, Italy. [3] Centre for Atmospheric Science, Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK. [4] Cooperation Group comprehensive molecular analytics/Joint Mass Spectrometry Centre, Helmholtz Zentrum München, 85764 Neuherberg, Germany. [5] Chair of Analytical Chemistry/Joint Mass Spectrometry Centre, Institute of Chemistry, University of Rostock, 18051 Rostock, Germany. [6] Aix Marseille Université, CNRS, LCE FRE 3416, 13331 Marseille, France. † Present addresses: INRA, UMR Eco and Sols, 2 Place Pierre Viala, 34060 Montpellier, France (M.C.); Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), UTAPRAD-DIM, Via E. Fermi 45, 00044 Frascati, Italy (R.C.). Correspondence and requests for materials should be addressed to A.S.H.P. (email: andre.prevot@psi.ch).

Particulate matter (PM) damages health[1] and affects climate[2]. Road vehicles are a significant source of PM, particularly in urban areas. A number of recent studies have shown that a large fraction, possibly the largest, of vehicular PM is secondary organic aerosol (SOA) produced via atmospheric oxidation of precursor gases in the exhaust[3–5]. Thus, understanding vehicular air pollution requires an assessment of SOA formation from different vehicle types. Two-stroke (2S) scooters (powered two-wheeled vehicles with engine displacement $\leq 50\,cm^3$) are popular globally, particularly in Asia, Africa and Southern Europe. Despite being high emitters of primary PM[6,7], regulations for scooters are generally less stringent than for other vehicles, for example, in Europe having reached Euro 5/V (a fifth tranche of regulations), for passenger cars and trucks, versus only Euro 2 for scooters (see Supplementary Table 1 and ref. 8). Accordingly, a scientific report to the European Commission suggests that scooters will emit more volatile organic compounds (VOCs) than all other vehicles combined in Europe by 2020 (ref. 9). Furthermore, high PM levels and toxic aromatic hydrocarbons, important SOA precursors[10], have been observed in many cities, especially in Asia[11]. Globally, organic aerosol (OA) dominates PM, with SOA accounting for the largest fraction[4].

Here we show that 2S scooters emit significant amounts of primary organic aerosol (POA), aromatic VOCs and also produce significant SOA. We use the term 'asymmetric polluter' to describe these vehicles as their emission factors (EFs) and evidence from air quality measurements before and after bans on scooters in Asian cities suggest they may dominate vehicular pollution despite their relatively small numbers. Chemical analysis of the emissions shows that SOA is mainly produced via photo-oxidation of aromatic VOCs, present in gasoline, from the exhaust. This shows that the known issue of incomplete fuel combustion during the 2S cycle is also responsible for SOA formation. Finally, we present the first online measurements of aged exhaust showing that SOA formation also produces reactive oxygen species (ROS) with potentially detrimental effects on our lungs.

Results

Emission Factors. We investigated POA emissions and SOA formation from 2S scooters and their potential health effects. The oxidation of VOCs in 2S scooter emissions produces significant SOA (g carbon (C) kg^{-1} fuel), with total OA on average 2.9 and 2.4 times higher than POA after aging for idling and driving 2S scooters, respectively (Fig. 1, and Supplementary Table 2). In addition, substantial toxic aromatic emissions (up to ~40% of emitted VOC volume for the scooters of this study) of benzene, toluene and C2–C4 alkylated benzenes, which are recognized SOA precursors[10,12], are present in the exhaust. Among the aromatics, benzene is of particular concern due to its carcinogenicity. Levels in the raw 2S scooter exhaust were as high as $300{,}000\,\mu g\,m^{-3}$ or 146 p.p.m.(v) from idling. The EU annual mean limit for the protection of human health is $5\,\mu g\,m^{-3}$ (ref. 13), while the US National Institute for

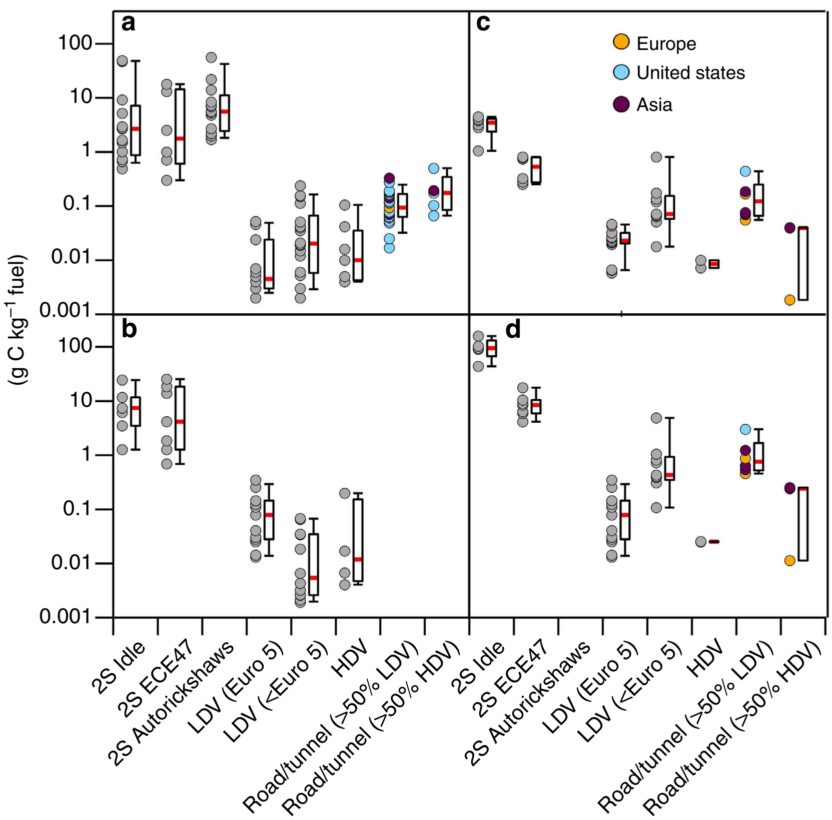


Figure 1 | Emission factors from scooters and other vehicles. EFs plotted as box-and-whiskers (median line, red; 25th and 75th percentile, box; 10th and 90th percentile, whiskers) of (**a**) POA, (**b**) aged OA (POA + SOA formation), (**c**) benzene and (**d**) light aromatics (benzene, toluene and C2–C4 alkylated benzenes). Points shown next to the box-and-whiskers are the individual data points, coloured depending on measurement region for ambient data. 2S scooters (this study, $n = 3$) were run in idle or during driving cycles (ECE47). Data on the other vehicles shown are from the literature (Supplementary Table 3) for light-duty and heavy-duty vehicles (LDVs and HDVs). LDVs data are further divided between vehicles meeting Euro 5 and those not meeting Euro 5, labelled <Euro 5 in parenthesis. Ambient data are split according to a contribution of HDVs to the data of higher than or lower than 50%. Note that many of the higher ambient values are from older vehicle studies (Supplementary Table 3).

Occupational Safety and Health recommends that workers wear special breathing equipment when exposed to benzene at levels exceeding 1 p.p.m. for 15 min. Waiting in traffic behind a 2S scooter, for example, at junctions and while the scooter is idling, may therefore be highly deleterious to health.

Secondary organic aerosol yields. The contribution of the aromatics to SOA formation was estimated by calculating an apparent aerosol yield, $y_{apparent}$, assuming all SOA comes from aromatic precursors:

$$y_{apparent} = C_{SOA} / \sum_{i} \Delta_i \quad (1)$$

where C_{SOA} is the SOA produced ($\mu g\, m^{-3}$) for a given mass change in aromatic i (Δ_i, i = benzene, toluene or C2–C4 alkylated benzenes). Apparent yields closely match average concentration-weighted literature aromatic SOA yields[10] (Fig. 2a, Supplementary Note 1) for idling, complete ECE47 driving cycles and ECE47 phase one (Ph1), indicating that most SOA is from aromatic precursors (Fig. 2a). SOA from ECE47 phase two (Ph2) alone is underestimated by equation (1), suggesting SOA production from unidentified compounds, emitted by the hot engine. Note that the total emission during a full cycle is dominated by Ph1, that is, by cold engine emissions. Furthermore, a 'Van Krevelen diagram' illustrates the aging of 2S scooter emissions, from oxygen to carbon (O:C) ~ 0 to O:C ~ 0.6. This elemental composition is consistent with that of previously observed SOA from aromatic precursors[14] (Fig. 2b). We therefore conclude that SOA formation from 2S scooter emissions is likely from the oxidation of aromatics, in contrast to diesel SOA, which is predominantly from other precursors[15].

Comparison to other vehicle types and ambient data. Figure 1 also shows laboratory and ambient measurements of POA, light aromatic and benzene EFs from passenger cars and trucks (Supplementary Table 3). Ambient data are from roadside/tunnel measurements in the US, EU and Asia, and are split according to the fraction of light-duty and heavy-duty vehicles (LDVs and HDVs, respectively) at the measurement site. Note that the general trend is for lower EFs in newer studies, consistent with improvements in emission controls. Also shown are data from Indian in-use 2S autorickshaws for comparison to the European scooters of this study. Caution is required in such a comparison; however, although similar (both have 2S engines), these are a different vehicle class and were furthermore tested during a different driving cycle. In general, ambient EFs from Asian vehicles are in the same range as European and US vehicles, while emissions from in-use 2S rickshaws are slightly higher than from the European scooters of this study. POA emissions from 2S scooters are on average around 20 (maximum 2,780) times higher than ambient (light-duty dominant) values, and aged OA emissions on an average 53–771 times higher than laboratory studies on other vehicle types. It should be noted that absolute aerosol concentrations can influence EFs: higher measurement concentrations would lead to higher EFs[15]. SOA formation is most significant from idling scooter emissions, while smaller at higher engine loads. However, POA emissions are higher under the latter conditions, and the aggregate POA + SOA emission at high load is comparable with that from idling.

Reactive oxygen species. We also examined the health implications of the 2S scooter SOA (other than those from the mass increase) using online measurements of particle-bound, water soluble reactive oxygen species (ROS)[16], which are linked to negative health effects[17]. ROS are undetectable in POA, but accounts for 0.5–1% carbon in the aged OA, suggesting that PM emissions initially become increasingly toxic with aging (Fig. 3). Increasing ROS is consistent with the increased O:C ratio of the aerosol and in line with a previous study showing increased oxidative potential with aging for 2S scooter emissions, albeit at aerosol and oxidant loadings much higher than under ambient conditions[18]. After 1–2 h of irradiation, ROS stabilizes or decreases, as reported previously for organic peroxides, likely due to decomposition processes[19,20].

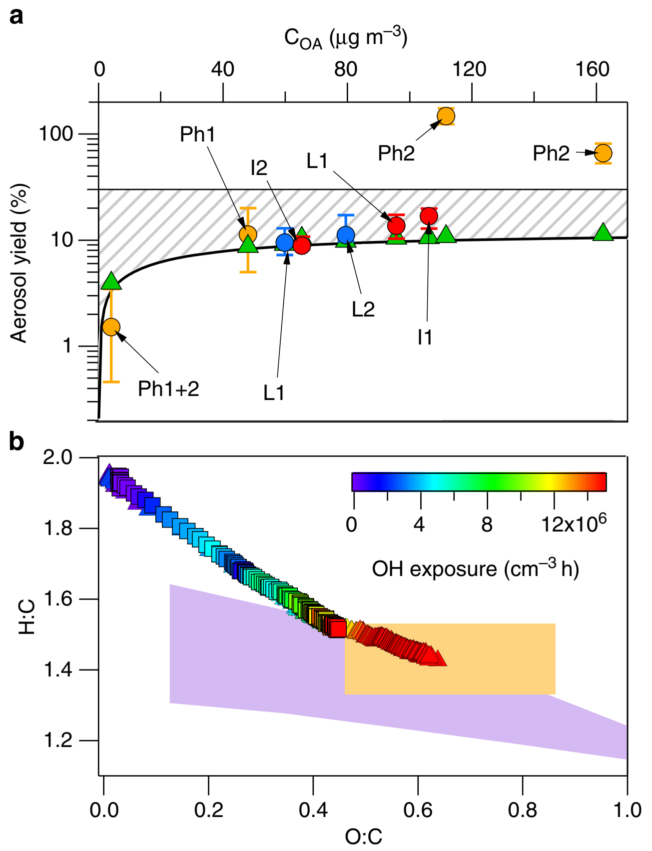


Figure 2 | Contribution of aromatic oxidation to two-stroke scooter secondary aerosol formation. (**a**) Apparent SOA mass yields, $y_{apparent}$ (equation (1)), as a function of suspended OA concentration (C_{OA}). Error bars show the sensitivity of $y_{apparent}$ to the chamber wall-loss factor, ± one s.d. $y_{apparent}$ for a Euro 1 and two Euro 2 2S scooters are shown in red, blue and orange, respectively. Ph 1 and Ph 2 are the first and second phases of the ECE47 driving cycle, I and L refer to idling and simulated low power, respectively. A predicted yield, concentration-weighted, for the mixture of all aromatics (Supplementary Note 1), is given in green triangles. The shaded region denotes the range between maximum (low NO_X) and minimum (high NO_X) SOA yields for *m*-xylene, a major aromatic constituent of gasoline. (**b**) Elemental ratios of OA emissions for the Euro 1 (squares) and a Euro 2 (triangles) scooter as a function of photochemical age. Elemental ratios observed for xylene[13] and ambient[31] SOA are shown, orange and purple, respectively.

Discussion

There are likely several reasons for these relatively large OA and aromatic emissions from 2S scooters. First, 2S engines, unlike four-stroke (4S), require addition of lubricant oils to the fuel, some of which is emitted in the exhaust. Second, during the 2S engine cycle some of the fresh fuel/air mixture passes directly through the engine[21], increasing VOC emissions, which may

explain the high SOA formation. Third, scooters generally utilize 'rich combustion' (low-air/fuel ratio), improving drivability while producing higher CO, VOC and PM emissions (but lower NO_X). Accordingly, the VOC emissions measured here, in particular aromatics as found in raw gasoline, are also on average 124 and 11 times higher from idling and driving 2S scooters, respectively,

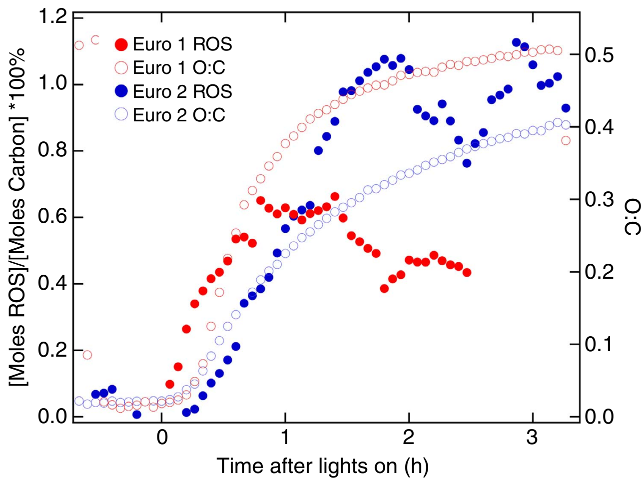


Figure 3 | Reactive oxygen species in two-stroke scooter emissions. The percentage of water soluble reactive oxygen species (ROS) and elemental O:C ratios of organic aerosol as a function of time after lights on in the smog chamber from Euro 1 (red) and Euro 2 (blue) 2S scooter exhaust emissions. ROS concentration measured in moles hydrogen peroxide equivalents is normalized to the molar organic carbon concentration per m^3 inside the smog chamber to give a percentage.

compared with those from other vehicles. Finally, scooter after-treatment systems are inherently inefficient due to their relatively small size and longer light-off times.

Precise estimation of a relative contribution to vehicular PM and aromatics from 2S scooters is difficult, since vehicle regulations vary by country. Another complication arises from the possibility of large contributions to OA from a small number of super-polluting vehicles (of all types). However, many scooters will likely fall into this super-polluting category, especially as a considerable number of scooters are in operation in some regions without any form of emissions control (note that all scooters presented in this study are equipped with two-way oxidation catalysts, which reduce emissions of carbon monoxide and VOCs) and because emissions may be further exacerbated by poor maintenance and tampering, rife for scooters[22]. Furthermore, ambient data in Fig. 1 likely include a number of such super-polluting vehicles. Therefore, our results suggest that 2S scooters are 'asymmetric polluters' of OA and aromatics compared with other vehicles. Using the average 2S scooter EF (ECE47 driving cycle) in Fig. 1 suggests that 2S scooters contribute to around 60% of roadside POA in Bangkok, where they account for 10% of fuel consumption (Fig. 4). In a more extreme case (comparing the 75th percentile for scooters and 25th percentile for ambient light-duty dominated), 2S scooters would contribute over 96% to roadside POA. Note that these values are based on the European scooters of this study. As Fig. 1 shows, emissions from some in-use Asian 2S vehicles may be higher, by a factor of three. Since other Asian vehicles are not expected to be more polluting based on Fig. 1, higher emissions from in-use Asian 2S vehicles would strengthen our conclusion that 2S scooters dominate urban pollution in the region.

Estimation of contributions to aged OA is more difficult as vehicular SOA has not been systematically quantified under

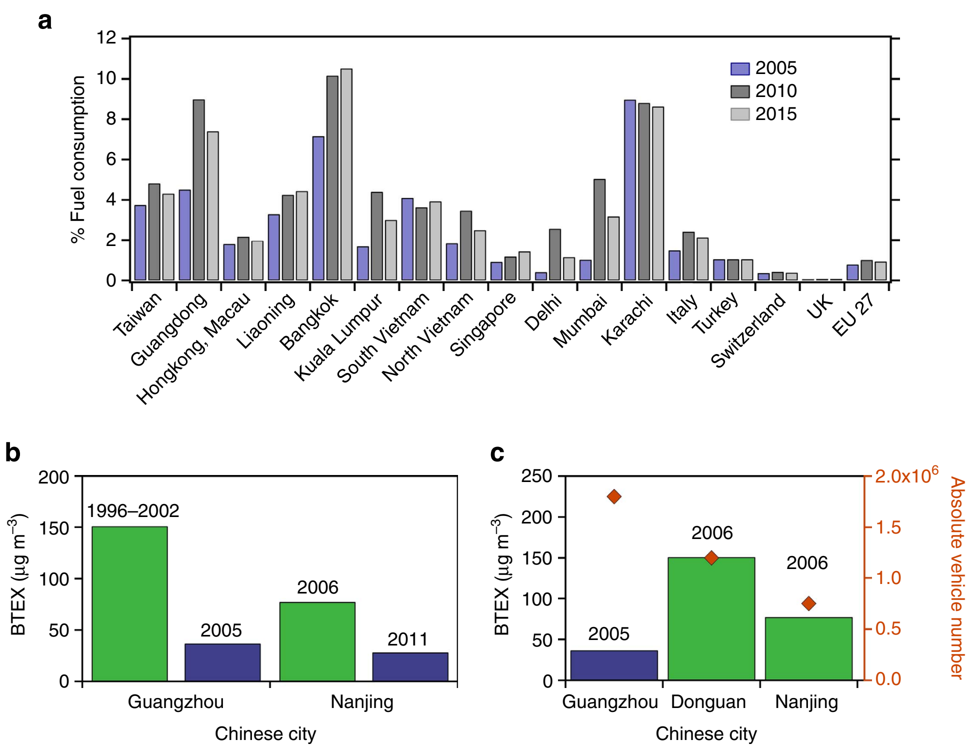


Figure 4 | Ambient and model data on two-stroke scooters. (**a**) Share of total fuel consumption by 2S scooters in 2005, 2010 and 2015 from the Greenhouse gas Air pollution Interactions and Synergies, GAINS, model[31]. (**b**) Roadside benzene, toluene, ethyl-benzene and xylene (BTEX) before (green) and after (blue) banning/restricting 2S scooters in two Chinese cities (**c**) Roadside BTEX and number of all vehicles in three Chinese cities, before (green) and after (blue) banning/restricting 2S scooters.

ambient conditions. However, smog chamber measurements suggest average-aged OA contributions to ambient vehicular PM of 85% (comparing with LDVs meeting Euro 5) or 98% (comparing with LDVs meeting less than Euro 5) from 2S scooters. Meanwhile, in the EU, 2S scooters consume only 1% of vehicle fuel, Fig. 4. Even with these low numbers, scooters may be the major source of some of the vehicle-related pollutants, especially in Southern Europe, and our data suggest that reducing the numbers of these vehicles would cost-effectively mitigate vehicle OA and aromatic emissions, given the alternatives available (electric and 4S). In this regard China has taken the lead, banning or restricting scooters in many cities since the late 1990s[23], leading to large decreases in the traffic-related aromatic emissions in some Chinese cities (Fig. 4b). Strikingly, roadside aromatics are now higher in Dongguan, where scooters are not banned, than 60 km away in Guangzhou, even though the traffic volume is much higher in Guangzhou (Fig. 4c). This result is statistically significant: year-to-year benzene, toluene, ethyl-benzene and xylene concentrations in Guangzhou were 229 $\mu g\,m^{-3}$ in 1996, 244 $\mu g\,m^{-3}$ in 1999, 290 $\mu g\,m^{-3}$ in 2000 and 150 $\mu g\,m^{-3}$ in 2002, average 228 ± 68 versus 37 $\mu g\,m^{-3}$ after the scooter ban in 2005, for example. Benzene, toluene, ethyl-benzene and xylene concentrations for Guangzhou and other cities reported in the literature are given in Supplementary Table 4.

Our data suggest that 2S scooters are a significant, and in many cities the largest, source of vehicular PM and toxic SOA and aromatic hydrocarbons, despite being a relatively small fraction of the total fleet. Therefore, given the alternative technologies available, restrictions on 2S scooters, already implemented in China, could improve air quality in many cities around the globe.

Methods

Measurement campaigns. We combine results from two measurement campaigns where 2S scooter exhaust was injected through a heated inlet into smog chambers[3,24,25] to produce SOA via photochemistry. During the first study, an in-use Euro 1 (E1) and a new Euro 2 (E2a) 2S scooter were run in idle or simulated low power. During the second campaign, emissions from a different Euro 2 2S scooter (E2b) were sampled during ECE47 driving cycles[26]. Supplementary Table 5 provides specifications of these vehicles. European (exhaust) emission standards are shown in Supplementary Table 1. Average OH concentrations were $\sim 5 \times 10^6\,cm^{-3}$. OH concentrations were determined from the decay of a nine times deuterated butanol (butanol-D9, 98% Aldrich) tracer as measured using a quadrupole proton transfer reaction mass spectrometer (idling 2S scooters) or proton transfer reaction time-of-flight mass spectrometer (Ionicon Analytik, driving cycle 2S scooters), see also ref. 27.

Estimation of the NO_X regime during experiments. Experiments were under 'high NO_X' conditions, which we define as the chemical regime where the main reactions of peroxy radicals (RO_2) are with NO rather than other peroxy radicals (self-reaction, or reaction with hydroperoxy radicals). An estimate of the ratio of the RO-forming reactions ($RO_2 + NO$) versus peroxide-forming reactions($RO_2 + RO_2$, $RO_2 + HO_2$) is possible for experiments conducted on idling scooters at the Paul Scherrer Institute chamber, instrumentation for which includes a NO_X monitor equipped with a 'blue light converter' (ensuring NO_X is truly measured as $NO_2 + NO$). Figure 5b shows the measured concentrations of NO and O_3, from an experiment conducted on 22 November 2010. This experiment was typical, with an initial VOC:NO_X ratio of around 50 and continuous addition of NO during photochemical aging.

The concentration of NO as a function of time t is given by:

$$\frac{d[NO]}{dt} = j_{NO2}[NO_2] - k_1[NO][O_3] - k_2[NO][RO_2 + HO_2], \quad (2)$$

Where j_{NO2} is the photolysis rate of NO_2 in the smog chamber ($0.01\,s^{-1}$) and k_1 ($1.8 \times 10^{-14}\,cm^3\,molecule^{-1}\,s^{-1}$) and k_2 ($7.7 \times 10^{-14}\,cm^3\,molecule^{-1}\,s^{-1}$) are the reaction rate constants for NO with O_3 and peroxides (CH_3O_2) at 298 K, respectively. Assuming a steady state of NO (only an approximation, Fig. 5 indicates this point is not reached until around 15 h of OH exposure (at OH = 10^6 molecules cm^{-3})), equation (2) can be written:

$$[RO_2 + HO_2] = \frac{j_{NO2}[NO_2] - k_1[NO][O_3]}{k_2[NO]}. \quad (3)$$

Equation (3) suggests NO concentrations at least an order of magnitude higher than $RO_2 + HO_2$ (for example, 14 times higher at OH = 10×10^6 molecules cm^{-3} h) during the experiment, based on concentrations measured inside the chamber (Fig. 5).

The branching ratio r between the RO_2/HO_2 reactions with NO versus other with peroxides (Fig. 5a) is determined using

$$r = \frac{k_2[NO]}{k_3[HO_2 + RO_2]}, \quad (4)$$

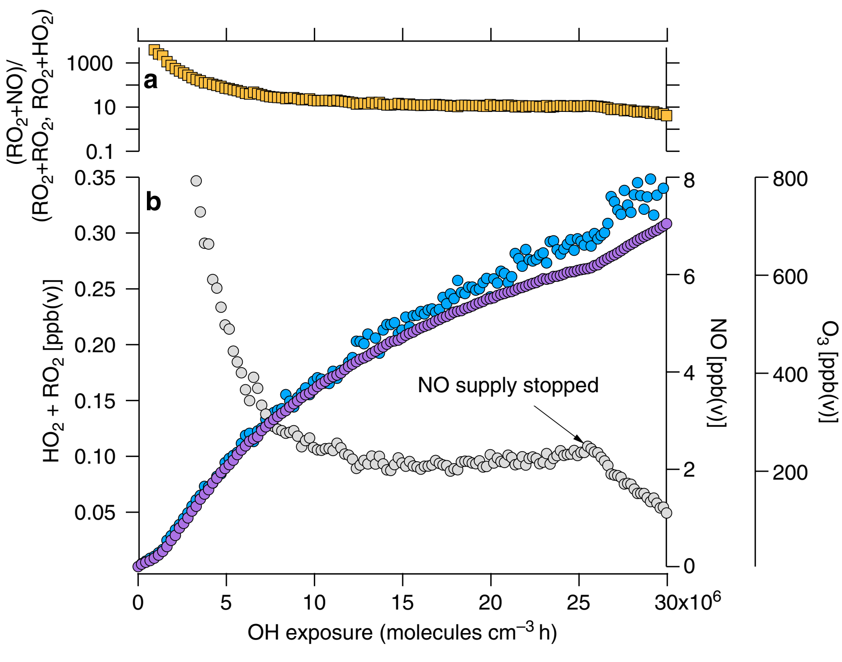


Figure 5 | Estimation of NO_X regime. (**a**) Calculated branching ratio between nitrate and peroxide reactions (orange squares) in the smog chamber during aging of emissions from an idling 2S scooter and (**b**) measured concentrations of ozone (O_3) (purple circles) and nitrogen monoxide (NO) (grey circles) as well as calculated peroxy radical ($HO_2 + RO_2$) (blue circles) concentrations.

where k_3 is the reaction rate constant between HO_2 and CH_3O_2 (7.7×10^{-12} cm^3 molecule^{-1} s^{-1}) at 298 K. We assume that the concentration of HO_2 and RO_2 is the same.

Figure 5a illustrates that the NO pathway is dominant, by at least a factor of 20 until OH $= 10 \times 10^6$ molecules cm^{-3} h, and by initially thousands of times higher, over the peroxide pathway. Since $r >> 1$ we consider these experiments 'high NO_X'.

An estimation of r during the tests on driving cycle emissions is complicated by the lack of an accurate NO_X instrument (that is, one equipped with a blue light converter). Furthermore, NO was only continuously added during the experiment where emissions were sampled from the cold phase (which featured the highest VOC:NO_X ratio of any experiment). However, given that the driving cycle tests generally produced higher NO emissions than the idling tests, and given that most fall on the yield curve in Fig. 2, we also assume $r >> 1$, although we can not rule out that during some experiments the conditions change from high to low NO_X. Although the VOC:NO_X ratios were high (around 50), our best estimate suggests that idling experiments were high NO_X throughout. Figure 5b shows the measured concentrations of NO and O_3, during the experiment.

Idling scooter experiments. Emissions were introduced into the 27 m^3 Paul Scherrer Institute Teflon environmental chamber[24]. The external temperature of the scooter exhaust was monitored (Thermocouple type K, Messelemente) and after an initial warming period of several minutes (consisting of idling or applying low power) the emissions were injected only when the external exhaust temperature was stable at idle or at simulated low power. Supplementary Table 2 provides the operating conditions, smog chamber OA concentrations and aerosol EFs of this study used in Fig. 1.

OA was monitored with a high-resolution time-of-flight aerosol mass spectrometer (Aerodyne). Unity collection efficiency is assumed, since emitted particles likely consist of spherical oil-like droplets with low bounce. After an initial spike in the OA concentration following sample injection, a time of at least 20 min was allowed for equilibration. The concentration of OA after this point was taken as the initial POA emission. A battery of 80 × 100 W ultraviolet black lights (ErgoLine 'Cleo Performance', Solarium lights), was used to initiate photo-oxidation and SOA formation. Experiments were carried out with a steady injection of NO (<20 ml min^{-1}) whereby NO was maintained at around 2–3 ppb(v). Relative humidity inside the smog chamber was between 40–60% for all experiments, and temperature was maintained at 25 °C.

OA was corrected for wall losses using

$$\mathrm{OA_{WLC}}(t) = \frac{\mathrm{OA_{MEAS}}(t)}{\exp(-kt)}, \quad (5)$$

where $\mathrm{OA_{WLC}}(t)$ and $\mathrm{OA_{MEAS}}(t)$ are the wall-loss corrected and measured organic matter concentrations, respectively, as a function of time t, and k is the first order mass-loss rate constant determined from an exponential fit of BC data.

VOCs inside the smog chamber were quantified with a quadrupole proton transfer reaction mass spectrometer, while carbon monoxide was quantified with a dedicated CO monitor (Aerolaser, CO-Monitor AL5002) and total gas-phase hydrocarbons were measured from the chamber using a flame ionization detector (FID, J.U.M model VE 7). Additional measurements at the tailpipe were performed by transferring emissions through a heated line (191 °C) to a Fourier transformed infrared spectrometer (MKS Multigas analyser 2030) for online measurements (at 1 Hz) of small hydrocarbons, nitrogen containing species (NO, NO_2, N_2O, NH_3 and HCN) and other oxygenated small organics (formaldehyde, acetaldehyde), as well as CO and CO_2.

Online reactive oxygen species measurements. Online particle-bound ROS analysis utilized the fluorescence probe 2,7-dichlorofluorescein in solution. Particles were collected and continuously extracted on a wetted hydrophilic filter. The particle collector samples air at 5 l min^{-1} and collects particles larger than aerodynamic diameter 50 nm with greater than 95% efficiency. Particles are collected and extracted in an aqueous solution of horseradish peroxidase (0.5 U ml^{-1}) allowing immediate reaction of ROS on collection. The concentration of ROS is characterized following subsequent reaction of the oxidized horseradish peroxidase with 2,7-dichlorofluorescein (5 μM) for 10 min at 40 °C, yielding the fluorescent product DCF in the continuous flow set up. The concentration of 2,7-dichlorofluorescein is measured using fluorescence spectroscopy in a flow-through cell and calibrated to ROS concentration with hydrogen peroxide. ROS data in Fig. 3 are normalized to the total carbon m^{-3}, determined from high-resolution fitting of aerosol mass spectrometer data, and presented as a percentage.

Driving cycle scooter experiments. The Paul Scherrer Institute mobile smog chamber[3] was deployed, and experiments conducted in a certified chassis dynamometer test cell (Vehicle Emissions Laboratories, Joint Research Centre of the European Commission, JRC-Ispra, Italy)[28,29]. Emissions from 2S scooters were sampled at the tailpipe during full ECE47 driving cycles, during Ph1 only of the ECE47 (first four modules of the driving cycle, Ph1), and during Ph2 only of the ECE47 (final four modules of the driving cycle, Ph2). The emissions were transferred to the smog chamber via a heated inlet system (150 °C) and Dekati ejector dilutor. Ultraviolet lights were switched on after several minutes to initiate photochemistry. OA concentrations were measured with a high-resolution time-of-flight aerosol mass spectrometer (Aerodyne), while black carbon was quantified with an aethalometer (AE33, Aerosol d.o.o.). The exponential decay rate of black carbon k was used in equation (5) to correct for particle losses to the walls. Gas-phase compounds were monitored with a proton transfer reaction time-of-flight mass spectrometer (Ionicon), while CO_2 and CO were measured using a cavity ring down spectrometer (Picarro, G2401) and total hydrocarbons were measured with a flame ionization detector (Horiba, THC Monitor APHA-370).

Emission factor determination. EFs from both measurement campaigns (EF, g C kg^{-1} fuel), (see also Supplementary Table 3), were calculated using a carbon mass balance:

$$\mathrm{EF} = \left(\frac{\mathrm{OC}}{C_{\mathrm{CO2}} + C_{\mathrm{CO}} + C_{\mathrm{HC}}}\right) \cdot W_{\mathrm{C}}, \quad (6)$$

where C denotes carbon mass, and the subscripts CO_2, CO, HC, carbon dioxide, carbon monoxide and hydrocarbon, respectively. W_c is the fuel carbon content (0.847 for gasoline).

For the idling scooter experiments, C_{CO} and C_{CO2} were measured at the tailpipe using the Fourier transformed infrared spectrometer. C_{HC} was measured from the smog chamber and scaled-up to the tailpipe concentration using the dilution ratio $\mathrm{CO_{tailpipe}/CO_{smog\ chamber}}$. Meanwhile, for the driving cycles all concentrations were determined at the smog chamber.

Emission factors from the literature. Figure 1 and Supplementary Table 3 show EFs calculated from the literature. When available, EFs are given as reported. OA EFs measured in tunnels/roadside are assumed to consist purely of POA and are converted to EFs in units of g kg^{-1} fuel using an organic matter to organic carbon ratio (OM:OC) of 1.2 (ref. 30). EFs given in units of g km^{-1} are converted using the following fuel consumptions (km kg^{-1}): Asia LDVs: 16.43; US LDVs: 14.93; EU LDVs: 18.20; Heavy-duty vehicles: 2.85. EFs measured during the Kansas City vehicle study are estimated by inserting per km EFs into equation (6). SOA in Supplementary Table 3 (g C kg^{-1} fuel) is converted from units of g kg^{-1} fuel using an OM:OC ratio of 2.0 (ref. 30). SOA formation from 2S scooters is not available in the literature, but was estimated from emissions of aromatic hydrocarbons using a yield (see Supplementary Note 1 and Fig. 2) of 8.4% (suspended OA concentration 50 μg m^{-3}). Further notes to individual studies are also provided in Supplementary Table 3.

References

1. Dockery, D. W. *et al.* An association between air pollution and mortality in six US cities. *New Engl. J. Med.* **329,** 1753–1759 (1993).
2. IPCC. *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2007).
3. Platt, S. M. *et al.* Secondary organic aerosol formation from gasoline vehicle emissions in a new mobile environmental reaction chamber. *Atmos. Chem. Phys.* **13,** 9141–9158 (2013).
4. Hallquist, M. *et al.* The formation, properties and impact of secondary organic aerosol: current and emerging issues. *Atmos. Chem. Phys.* **9,** 5155–5236 (2009).
5. Nordin, E. Z. *et al.* Secondary organic aerosol formation from idling gasoline passenger vehicle emissions investigated in a smog chamber. *Atmos. Chem. Phys.* **13,** 6101–6116 (2013).
6. Czerwinski, J., Comte, P. & Reutimann, F. Nanoparticle Emissions of a DI 2-Stroke Scooter with Varying Oil and Fuel Quality. *SAE transactions* **114,** 541–556 (2005).
7. Rijkeboer, R., Bremmers, D., Samaras, Z. & Ntziachristos, L. Particulate matter regulation for two-stroke two wheelers: necessity or haphazard legislation? *Atmos. Environ.* **39,** 2483–2490 (2005).
8. European Commission (EC). *O.J.E.C.* **L67,** 14–30 (2002).
9. Geivanidis, S. *et al. Arisotle University* (2008). *Thessaloniki,* available at http://www.ec.europa.eu/enterprise/sectors/automotive/documents/calls-for-tender-and-studies/index_en.htm, last accessed 12 January 2014.
10. Ng, N. L. *et al.* Secondary organic aerosol formation from m-xylene, toluene, and benzene. *Atmos. Chem. Phys.* **7,** 3909–3922 (2007).
11. Lan, T. T. N. & Minh, P. A. BTEX pollution caused by motorcycles in the megacity of HoChiMinh. *J. Environ. Sci.* **25,** 348–356 (2013).
12. Odum, J. R. *et al.* Aromatics, reformulated gasoline, and atmospheric organic aerosol formation. *Environ. Sci. Technol.* **31,** 1890–1897 (1997).
13. European Commission (EC), *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe,* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF (2008).
14. Chhabra, P. S., Flagan, R. C. & Seinfeld, J. H. Elemental analysis of chamber organic aerosol using an Aerodyne high-resolution aerosol mass spectrometer. *Atmos. Chem. Phys.* **10,** 4111–4131 (2010).
15. Robinson, A. L. *et al.* Rethinking organic aerosols: semivolatile emissions and photochemical aging. *Science* **315,** 1259–1262 (2007).

16. Fuller, S. J., Wragg, F. P. H., Nutter, J. & Kalberer, M. Comparison of on-line and off-line methods to quantify reactive oxygen species (ROS) in atmospheric aerosols. *Atmos. Environ* **92,** 97–103 (2014).
17. Donaldson, K. *et al.* Oxidative stress and calcium signaling in the adverse effects of environmental particles (PM_{10}). *Free Radic. Biol. Med.* **34,** 1369–1382 (2003).
18. McWhinney, R. D., Gao, S. S., Zhou, S. & Abbatt, J. P. Evaluation of the effects of ozone oxidation on redox-cycling activity of two-stroke engine exhaust particles. *Environ. Sci. Technol.* **45,** 2131–2136 (2005).
19. Mertes, P., Pfaffenberger, L., Dommen, J., Kalberer, M. & Baltensperger, U. Development of a sensitive long path absorption photometer to quantify peroxides in aerosol particles (Peroxide-LOPAP). *Atmos. Meas. Tech.* **5,** 2339–2348 (2012).
20. Surratt, J. D. *et al.* Chemical composition of secondary organic aerosol formed from the photooxidation of isoprene. *J. Phys. Chem. A* **110,** 9665–9690 (2006).
21. Spezzano, P., Picini, P. & Cataldi, D. Contribution of unburned lubricating oil and gasoline-derived n-alkanes to particulate emission from non-catalyst and catalyst-equipped two-stroke mopeds operated with synthetic lubricating oil. *J. Environ. Monit.* **10,** 1202–1210 (2008).
22. Santino, D., Picini, P. & Martino, L. Particulate matter emissions from two-stroke mopeds. *SAE* **2014,** 4–10 (2001).
23. Yang, C. J. Launching strategy for electric vehicles: Lessons from China and Taiwan. *Technol. Forecast. Soc. Change* **77,** 831–834 (2010).
24. Paulsen, D. *et al.* Secondary organic aerosol formation by irradiation of 1, 3, 5-trimethylbenzene-NOx-H_2O in a new reaction chamber for atmospheric chemistry and physics. *Environ. Sci. Technol.* **39,** 2668–2678 (2005).
25. Chirico, R. *et al.* Impact of aftertreatment devices on primary emissions and secondary organic aerosol formation potential from in-use diesel vehicles: results from smog chamber experiments. *Atmos. Chem. Phys.* **10,** 11545–11563 (2010).
26. United Kingdom Transport Research Laboratory (TRL), *A Reference Book of Driving Cycles for Use in the Measurement of Road Vehicle Emissions*, http://www.trl.co.uk/online_store/reports_publications/trl_reports/cat_traffic_and_the_environment/report_a_reference_book_of_driving_cycles_for_use_in_the_measurement_of_road_vehicle_emissions.htm (2009).
27. Barmet, P. *et al.* OH clock determination by proton transfer reaction mass spectrometry at an environmental chamber. *Atmos. Meas. Tech.* **5,** 647–656 (2012).
28. Clairotte, M. *et al.* Online characterization of regulated and unregulated gaseous and particulate exhaust emission from two-stroke mopeds: a chemometric approach. *Anal. Chim. Acta.* **717,** 28–38 (2012).
29. Adam, T. *et al.* Chemical characterization of emissions from modern two-stroke mopeds complying with legislative regulation in Europe (EURO-2). *Environ. Sci. Technol.* **44,** 505–512 (2010).
30. Aiken, A. C. *et al.* O/C and OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-flight aerosol mass spectrometry. *Environ. Sci. Technol.* **42,** 4478–4485 (2008).
31. Klaassen, G., Berglund, C. & Wagner, F. *The GAINS Model for Greenhouse Gases-Version 1.0: Carbon Dioxide (CO_2).* IIASA Interim Report IR-05-53 (International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, 2005).

Acknowledgements
This work was supported by the Swiss Federal Office for the Environment (FOEN), the Federal Roads Office (FEDRO), the Swiss National Science Foundation (Ambizione PZ00P2_131673, SAPMAV 200021_13016), the EU commission (FP7, COFUND: PSI-Fellow, grant agreement n.° 290605), the UK Natural Environment Research Council (NERC), the French Environment and Energy Management Agency (ADEME, Grant number 1162C00O2) and the Velux Stiftung, project 593.

Author contributions
Study design (S.M.P., A.S.H.P., M.C., A.A.Z., U.B., I.E.H.); experimental work, idling 2S scooters (S.M.P., M.C., L.P., P.B., J.D., S.J.F.); experimental work, driving cycle 2S scooters and other vehicles (S.M.P., I.E.H., J.G.S., M.C., A.A.Z., S.H., B.T.-R. S.M.P., R.S.-B.); data analysis, emission factors (S.M.P., S.M.P., I.E.H., M.C., A.A.Z., J.G.S., P.B., S.H., B.T.-R.); data analysis, reactive oxygen species (S.J.F.); data analysis, SOA yields (S.M.P.); literature data ambient BTEX (R.-J.H.); literature data ambient PM/aromatic (I.E.H.); writing of manuscript (S.M.P.); preparation of display items (S.M.P., S.M.P., R.-J.H.); data interpretation and editing of manuscript (S.M.P., M.C., I.E.H., S.M.P., R.-J.H., R.Z., R.C., C.A., J.D., L.P., A.S.H.P., M.K., N.M., U.B.); comments and discussion on the manuscript (all).

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Platt, S. M. *et al.* Two-stroke scooters are a dominant source of air pollution in many cities. *Nat. Commun.* 5:3749 doi: 10.1038/ncomms4749 (2014).

nature COMMUNICATIONS

ARTICLE

Received 4 Dec 2012 | Accepted 30 Apr 2013 | Published 4 Jun 2013

DOI: 10.1038/ncomms2961

Urban characteristics attributable to density-driven tie formation

Wei Pan[1], Gourab Ghoshal[1,†], Coco Krumme[1], Manuel Cebrian[1,2,3] & Alex Pentland[1]

Motivated by empirical evidence on the interplay between geography, population density and societal interaction, we propose a generative process for the evolution of social structure in cities. Our analytical and simulation results predict both super-linear scaling of social-tie density and information contagion as a function of the population. Here we demonstrate that our model provides a robust and accurate fit for the dependency of city characteristics with city-size, ranging from individual-level dyadic interactions (number of acquaintances, volume of communication) to population level variables (contagious disease rates, patenting activity, economic productivity and crime) without the need to appeal to heterogeneity, modularity, specialization or hierarchy.

[1] Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2] Department of Computer Science and Engineering, University of California at San Diego, La Jolla, California 92093, USA. [3] National Information and Communications Technology Australia, Melbourne, Victoria 3010, Australia. † Present address: Department of Earth & Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. Correspondence and requests for materials should be addressed to A.P. (email: pentland@mit.edu).

A larger percentage of people live in cities than at any point in human history[1], while the density of urban areas is generally increasing[2]. One of the enduring paradoxes of urban economics concerns why people continue to move to cities, despite elevated levels of crime, pollution, disease and wage premiums that have steadily lost ground to premiums on rent[3]. New York in the 18th century, according to Thomas Jefferson, was 'a toilet of all the depravities of human nature'. Since Jefferson's day, the city has grown to host the depravities of 100-fold more people, yet the stream of new arrivals has not stemmed.

While the forces behind any urban migration are complex, the advantages afforded by urban density comprise an important driver. Smith[4] was one of the first to point to urban centres as exceptional aggregators, whether of innovations or depravities. Cities appear to support levels of enterprise impossible in the countryside, and urban areas use resources more efficiently, producing more patents and inventions with fewer roads and services per capita than rural areas[5–10].

Despite the widespread focus on density as a driver for the uniqueness of cities both in scientific and popular audiences, we still lack a compelling generative model for why an agglomeration of people might confer an advantage. Important advances in several fronts have highlighted the difficulty in gaining an understanding of the urban processes beyond the density description level. Early economic models of agglomeration point to the role of technology diffusion in creating intellectual capital[11–13], but lack a quantitative description of the generative mechanism for how this diffusion happens. Hierarchies have also been proposed as an elegant mechanism for this growth[14]; however, recent studies hint at the absence of well-defined hierarchy across geographical scales[15–19]. It has also been observed[20] that diversity among residents and their intermingling displays a weak correlation with cities' success thus prompting the authors to conclude 'more fine-scale data on interactions among people of different disciplines—or the culture, laws and peculiarities of cities—is required to better assess the under- or overperformance of innovation of cities'.

Recent developments in the study of social networks shed some light on this challenge. Empirical evidence suggests that interactions and information exchange on social networks are often the driving force for idea-creation, productivity and individual prosperity. Examples of this include the theory of weak ties[21,22], structural holes[23], the strong effect of social interaction on economic and social success[24], the influence of face-to-face interactions on the effect of productivity, as well as the importance of information flow in the management of Research and Development[25,26]. Consequently, it seems that understanding the mechanism of tie formation in cities is the key to the development of a general theory for a city's growth described by it's economic indicators and its population. Following this line of thinking, our proposed answer for super-linear growth of cities can be regarded as a natural extension of Krugman's insights on industries[6]. Krugman pointed out the connection between manufacturing efficiency and transportation of goods as a function of proximity of factories. Similarly, our theory connects the efficiency of idea-creation and information flow to the proximity of individuals generating them.

These ideas have a long lineage in urban sociology, urban geography and economic geography. Louis Wirth, for example, conceptualized these processes in the late 1930s (ref. 27), prompting a vast literature in economic geography that explores the relationship between density and innovation, as well as that between diffusion (via social ties) and population density (going back to Hagerstrand in the 1950s (refs 28,29); and a more recent, but well-established literature on density and creativity (see, for example, Richard Florida's work[30–32]).

In this paper, we present a simple, bottom-up, robust model describing the efficient creation of ideas and increased productivity in cities. This contribution's goal is to integrate these ideas into a single mathematical model that can be tested against available empirical data. Our model consists of two essential features. We propose a simple analytical model for the number of social ties $T(\rho)$ formed between individuals, with population density ρ as its single parameter. We demonstrate that increases in density and proximity of populations in cities leads to a super-linear growth of *social-tie density* for urban population. We then show that the diffusion rate along these ties—a proxy for the amount and speed of information flow and idea adoption—accurately reproduces the empirically measured scaling of urban features such as rate of AIDS/HIV (acquired immunodeficiency syndrome/human immunodeficiency virus) infections, communication and GDP (gross domestic product). The model naturally leads to a super-linear scaling of indicators with city population[9] without the need to resort to any parameter tuning (although it predicts a different functional form than a simple power–law and is a more accurate match to the data). The surprisingly similar scaling exponent across many different urban indicators (see Supplementary Note 1 and Supplementary Table S1), suggests a common mechanism behind them. Social-tie density and information flow, therefore, offer a parsimonious, generative link between human communication patterns, human mobility patterns and the characteristics of urban economies, without the need to appeal to hierarchy, specialization or similar social constructs.

Results

A model for social-tie density. We propose to model the formation of ties between individuals (represented as nodes) at the resolution of urban centres. As our model is based on geography, a natural setting for it is a two-dimensional Euclidean space with nodes denoted by the coordinates $\vec{x}_i \in \mathbf{R}^2$ on the infinite plane. Furthermore, we also assume that these nodes are distributed uniformly in space, according to a density ρ defined as,

$$\rho = \text{no. of nodes per unit area.}$$

While the assumption of uniform density is an approximation, the qualitative features of the model are unaffected by other more realistic choices of the density distribution—see Supplementary Note 2 and Supplementary Fig. S1. Following Liben-Nowell *et al.*[33], we define the probability of a tie to form between two nodes i, j in the plane as

$$P_{\mathrm{ij}} \propto \frac{1}{\mathrm{rank_i(j)}}, \qquad (1)$$

where the rank is defined as

$$\mathrm{rank_i(j)} := |\{k : d(\mathrm{i},k) < d(\mathrm{i},\mathrm{j})\}|, \qquad (2)$$

and d_{ij} is the Euclidean distance between the two nodes. If j lies at a radial distance r from node i, then the number of neighbours closer to i than j is the product of the density and the area of the circle of radius r, and thus the rank is simply,

$$\mathrm{rank_i(j)} = \rho\pi r^2, \qquad (3)$$

which implies that the probability an individual forms a tie at distance r goes as $P(r) \sim 1/\pi r^2$, similar in spirit to a gravity model[34].

For a randomly chosen node, integrating over r up to an urban mobility 'boundary' denoted as $r_{\max}$, we obtain the expected number of social ties for this chosen node denoted as $t(\rho)$.

$$t(\rho) = \ln\rho + C, \qquad (4)$$

where $C = 2\ln r_{\max} + \ln\pi + 1$. We note that $r_{\max}$ may well be unique for each city, and is often determined by geographical

constraints, as well as city infrastructure (cf. Supplementary Note 3 and Supplementary Fig. S2–S3). Integrating over the number of social ties for all nodes within a unit area gives us the social-tie density $T(\rho)$,

$$T(\rho) = \rho \ln \rho + C' \rho, \tag{5}$$

with $C' = C - 1$. Thus the density of social ties formed between individuals grows as $T(\rho) \sim \rho \ln \rho$, a super-linear scaling consistent with the observations made by Calabrese *et al.*[35] (also discussed in the content below). We argue that $T(\rho)$ to a first approximation is the individual dyadic-level ingredient behind the empirically observed growth of city indicators. For more detail on the theoretical analysis and support for the assumptions involved, see Supplementary Note 4 and 5 and Supplementary Figs S4–S6.

In order to test this theoretical result, we perform simulations of tie formation with more realistic discrete settings. Urban areas differ dramatically in both regional boundaries and population density. It is thus important to test the sensitivity of the model to a diversity of input parameters for the density ρ and the urban 'boundary' $r_{\max}$. We start from an empty lattice of size $N \times N$, with N^2 possible locations. The density ρ is gradually increased by randomly assigning new nodes to empty locations on the grid, where each node represents a small community, or city block of 10^2 individuals. Once a node is added, the probability of forming a tie with one of its existing neighbours is computed by counting the number of nodes closer to this node according to equation (1). To test the sensitivity of our results to the relevant parameters we vary the size of the grid ($20 \leq N \leq 400$) of blocks to mimic different scales for city boundaries $r_{\max}$ assuming $r_{\max}$ is the size of the grid. In addition, we also vary city population between 10^4 and 10^7 residents, as well as the functional form of the density distribution.

In Fig. 1 we show the average over 30 realizations of the simulation for different values of the grid size N and city boundary $r_{\max}$. The density ρ in this case represents the relative percentage of occupied locations on the grid, and $T(\rho)$ the total number of ties formed between nodes. As Fig. 1 shows, the agreement between the theoretical expression for $T(\rho)$ equation (5) and the curves generated by the simulation, is excellent at all scales despite our continuum approximation ($R^2 \approx 1$).

As a comparative exercise, on the same plot, we also show the best fit to the form $T(\rho) \sim \rho^{\beta}$ and find a value of $\beta \approx 1.16$. We note that this value is strikingly similar to empirically observed values by fitting a power–law to the relationship

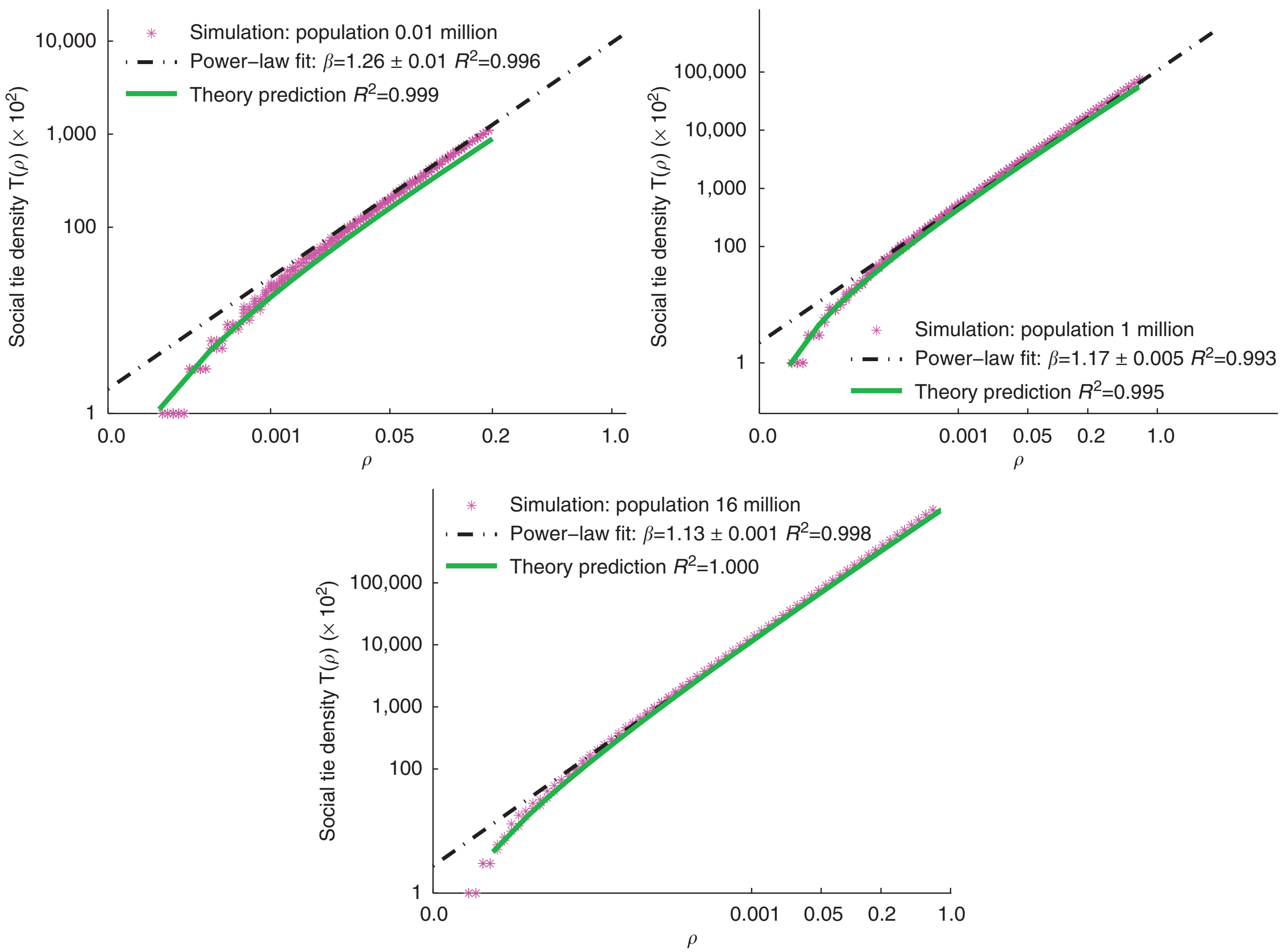


Figure 1 | The number of social ties as function of grid sizes and urban mobility limits. The number of ties $T(\rho)$ plotted as a function of ρ for various grid sizes N. The data points represent the average over $n = 30$ realizations of the simulation described in the text, while the solid green line is the theoretical expression equation (5). The dashed line is a fit to the form $T(\rho) \sim \rho^{\beta}$. As can be seen in each case the agreement between theory and simulation is excellent. The best fit to the scaling exponent yields a value of $\beta \approx 1.15$ independent of N. Note that the measured value of the exponent in empirical data is $1.1 \leq \beta \leq 1.3$.

between population and urban indicators. It has been suggested that a fit of the form $x \ln x$ can easily be mistaken for x^{β} (ref. 36), which together with our model suggests that the observed scaling of cities may alternatively be described by equation (5). The latter functional form is additionally supported by the fact that it represents a generative model for the emergence of urban features as a result of density-driven communication patterns, without any parameter tuning or a priori assumption about the structure of the underlying social network. Our simulation results indicate that the scaling described in equation (5) is robust with respect to the choice of different functional forms for the density distribution. (Supplementary Note 2 and Supplementary Fig. 1).

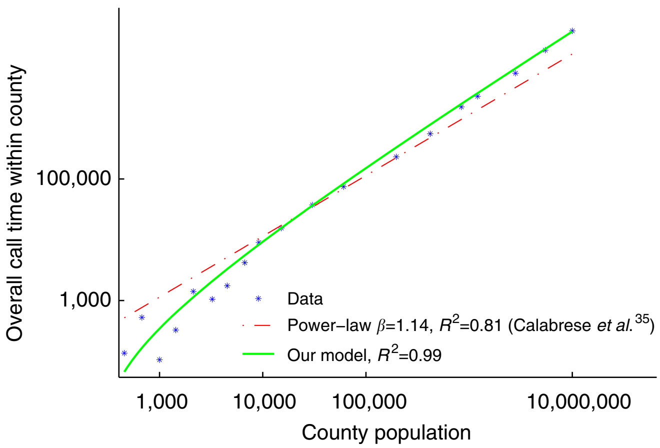


Figure 2 | Overall time of calls between residents of a county as a function of its population. The points refer to the data (adapted from Calabrese *et al.*[35] computed from ten million users' mobile phone call records within United States during July 2010), while the solid line is the theoretical prediction from the model equation (5) adapted to raw population. The model captures both the super-linear growth and tilts on both ends of the curve while providing a superior fit to the data (based on adjusted R^2-value) when compared with a pure power-law relation (dashed curve).

Empirical evidence for the effect of social-tie density. Recent work[35] shows a super-linear relationship between calling volume (time) and population across different counties in the United States. As Fig. 2 illustrates, the super-linear relationship in the data is approximated by the authors as a power–law growth $y = ax^{\beta}$ with $\beta \approx 1.14$. However, by assuming a uniform distribution on county sizes and treating population as a proxy for density, we show that our density-driven model is able to capture precisely the distribution of the call volume. The model produces the exact shape of the curve, including the power–law growth pattern ($\beta = 1.14$) and tilts on both end, with an adjusted $R^2 = 0.99$ (see Fig. 2). Consequently, we propose that the model may well provide a reasonable explanation for communication patterns observed in US counties.

Information diffusion and adoption with social-tie density. We note that the expected patterns of link and interaction formation in itself is insufficient to explain how growth processes in cities work to create observed certain scaling phenomena such as productivity and innovations. Instead, we believe that the manner in which these links spread information and encourage idea and behaviour adoptions actually determines value-creation and productivity. As it is known that social network structure has a dramatic effect on the access of information and ideas[21,24,23,25,26], it seems plausible that higher social-tie density should engender greater levels of idea spreading leading to the observed increases in productivity and innovation.

To test the hypothesis that a city's productivity is related to how far information travels and how fast its citizens gain access to innovations or information, it is natural to examine how this information flow scales with population density, and to quantify the functional relationship between link topology and speed of information spreading. We, therefore, simulated two models of contagion of information diffusion[37–39] on networks generated by our model. The first contagion model simulates diffusion of simple facts, where a single exposure is enough to guarantee transmission. The second more complex diffusion model is typical of behaviour adoption, where multiple exposures to a new influence/idea is required before an individual adopts it. In Fig. 3,

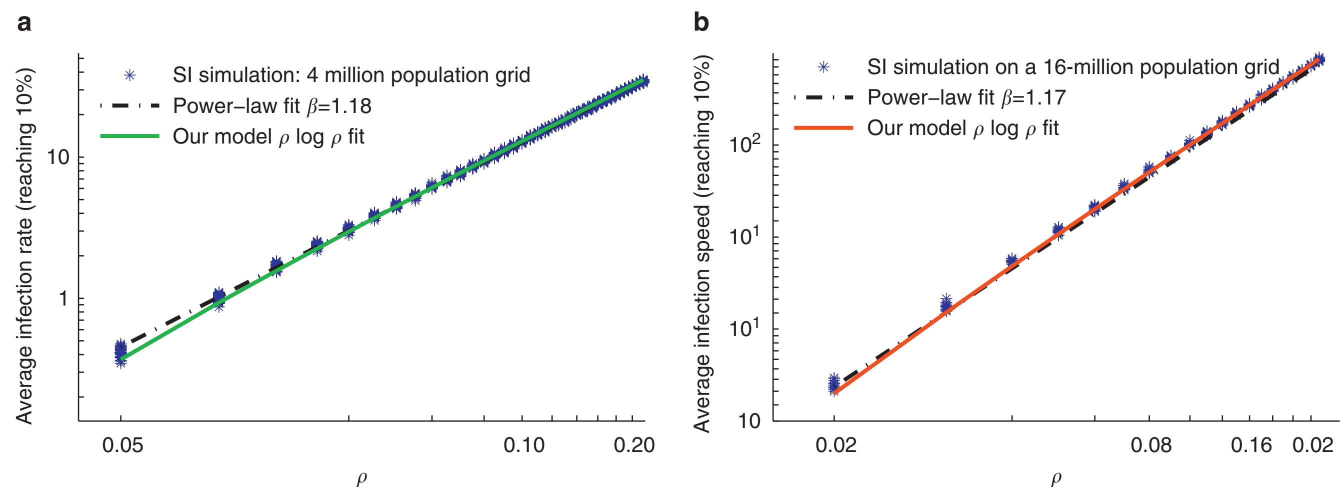


Figure 3 | The spreading rate as a function of density for two different contagion models. (**a**) The mean spreading rate as a function of density ρ. The points correspond to $n = 30$ realizations of simulations of the SI model on a 200×200 grid. The dashed line corresponds to a fit of the form $R(\rho) \sim \rho^{1+\alpha}$ with $\alpha = 0.18$. The solid line is a fit to the social-tie density model. (**b**) The mean spreading rate as a function of ρ under the complex contagion diffusion model based on $n = 30$ realizations of simulations. The dashed line corresponds to the power-law fit of the form $R(\rho) \sim \rho^{1+\alpha}$ with $\alpha = 0.17$. Once again the solid line is the fit to the model described in the paper. In both cases, the social-tie density model provides a better fit than a simple power-law with with much lower mean-square errors (29% and 41% lower respectively).

we discover that in both susceptible-infectious (SI) and complex contagion models the mean diffusion speed grows in a super-linear fashion with $\beta \approx 1.2$, in line with our previous results and match well with the disease spreading indicators in cities[9]. As a consequence we conclude that an explanation for the observed super-linear scaling in productivity with increasing population density is the super-linear scaling of information flow within the social network.

Population level variables. While in most cases it is not possible to obtain the social-tie density of a city directly, our model suggests that population density is strongly correlated with social-tie density across cities with similar transportation infrastructure and economic situations (that is, similar $r_{\max}$). Therefore, we here explore social-tie density indirectly by using population density measures, and we only focus on horizontal comparison of cities of similar levels of economic development, such as US cities and European Union cities.

As a test case for our hypothesis, we study the prevalence of AIDS/HIV infections in cities in the United States. In Fig. 4, we plot the prevalence of AIDS/HIV in 90 metropolitan areas in 2008 (ref. 40) as a function of population density. As the figure indicates, there is fairly good agreement between the data and the curve generated by our model of diffusion using both the simple and complex contagion models.

The same agreement holds for European cities on economic indicators. In Fig. 5, we plot the overall GDP per square km in NUST-2 (Nomenclature of Territorial Units for Statistics level-2) regions in the European Union as a function of population density ρ, as well as population size. The NUST-2 regions are defined by the European Union as the city-size level territorial partition for census and statistics purposes[41]. We find a strong positive correlation between density and the corresponding urban metric with a super-linear scaling component, but conversely a much weak and sublinear growth pattern on raw population size. While it is not the main focus of this paper, we show that the super-linear growth on density can be often be indicated in data

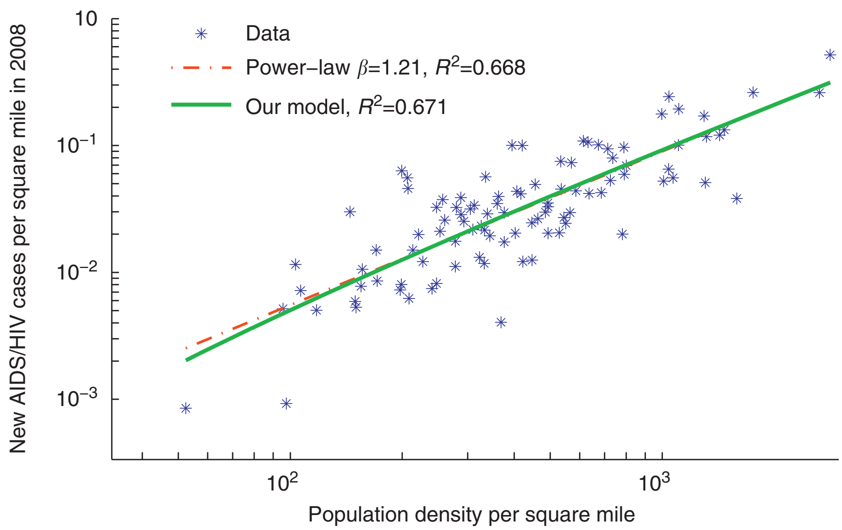


Figure 4 | Spreading rate of HIV as a function of density in US Metropolitan Statistical Areas. The relationship between density and AIDS/HIV spreading rate of the 90 metropolitan statistical areas from recent Centers for Disease Control and Prevention and US Census surveys. As is visible, the model captures the qualitative trends in the data.

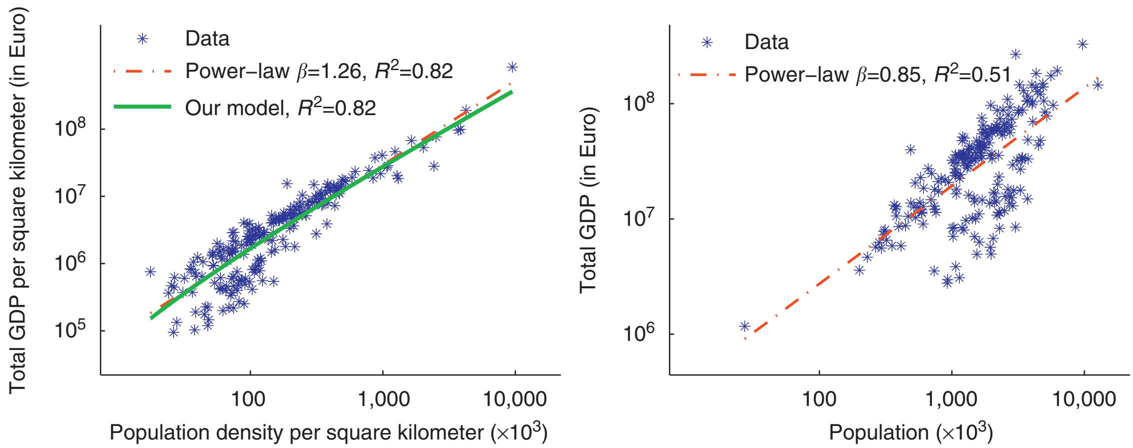


Figure 5 | Correlation between GDP and population, as well as correlation between GDP and population density for all 247 NUST-2 regions in the European Union. Left panel: correlation between density and GDP, suggesting a strong correlation with a super-linear functional form as predicted by the model. A pure power–law fit to the data is also shown for illustrative purposes. Right panel: the correlation between population and GDP this time showing a sublinear functional form. However, the poor R^2-value suggests that raw population does not correlate, as well as density with GDP growth in cities.

as super-linear growth on population, and that density is a better indicator for socio-economic growth than population—see Supplementary Note 4.

Note that in both data sets the scaling exponents are restricted within a narrow band $1.1 \leq \beta \leq 1.3$, potentially suggesting a common mechanism behind both the prevalence of AIDS/HIV and scaling of GDP with respect to the population density. An advantage afforded by our model is the need to dispense with parameter tuning, as the model naturally produces this scaling within a reasonable margin of error. Thus, by considering social structure and information/disease flow as a major driving force in many of the city indicators, our approach provides a unique and general theory to the super scaling phenomena of cities.

Both the spreading of information (potentially leading to increase productivity and innovation) and contagious diseases rely on the mechanism of social interactions. However, while information can mediate via Mass Media (exogenous) influence, and/or endogenous (word-of-mouth) processes, we chose to highlight the AIDS/HIV spreading data to validate our model, as an example of a purely endogenous process.

Discussion

In this paper, we propose social-tie density (the density of active social ties between city residents) as a key determinant behind the global social structure and flow of information between individuals. Based on this, we have described an empirically grounded generative model of social-tie density to account for the observed scaling behaviour of city indicators as a function of population density. Our model accurately explains how urban density drives the super-linear growth of social interaction density[35], and eventually the super-linear growth of productivity as observed in many empirical data sets.

The conceptual distinction between density and social density is an important one, as it has been missing in popular accounts that may percolate to urban planners and policy-makers. As a matter of fact, throughout the 20th century, the United States has witnessed major shifts from dense to suburban, then back to dense urban planning. As recently as 1970, suburban populations surpassed the urban one—motivated by a search for an idealized small, low-density, locally oriented community[42]. Indeed, these back and forth shifts may have been facilitated by an incomplete understanding of the benefits afforded by urban density.

The model predicts that social-tie density scales super-linearly with population density, while naturally accounting for the narrow band of scaling exponents empirically observed across multiple features and different geographies. We note that this is achieved without the need to resort to parameter tuning or assumptions about heterogeneity, modularity, social hierarchies, specialization or similar social constructs. We, therefore, suggest that population density, rather than population size per se, is at the root of the extraordinary nature of urban centres. As a single example, metropolitan Tokyo has roughly the same population as Siberia while showing remarkable variance in criminal profile, energy usage and economic productivity. We provide empirical evidence based on studies of indicators in European and American cities (both categories representing comparable economic development), demonstrating that density is a superior metric than population size in explaining various urban indicators.

We note that current technology makes remote communication and collaboration extremely easy and convenient; however, the importance of packing people physically close to each other is still widely emphasized[43–45]. We postulate that cities potentially operate under the same principle–as a consequence of proximity and easy face-to-face access between individuals—communication and ultimately productivity is greatly enhanced. Thus, though it is reasonable to surmise that individuals migrate to the city for reasons connected with individual needs and preferences, our argument suggests that, it is the benefits afforded by social-tie density that maintains them as residents.

While our model provides a fundamental *first-principles* basis for explaining productivity of cities, we note the importance of *higher-order* variables such as transportation infrastructure in order to tailor the model to specific cases to get better results. As an example, the density of social ties is intrinsically a function of the ease of access between residents living in the same city. Consider the case of Beijing, which has a very high population density, but due to its traffic jams, is currently *de-facto* divided into many smaller cities with limited transportation capacities between them. Consequently, it may not demonstrate a higher social-tie density than other cities with a much lower population density. Thus a direct comparison of the model predictions with a similarly dense area such as Manhattan needs to take into account this refinement. In keeping with the spirit of the simplicity and bottom-up approach of our model, we chose to use data from cities within the United States and the European Union such that extraneous variables are controlled for.

A number of theories of urban growth suggest the importance of specialist service industries, or high-value-add workers, as generative models of city development. While our model does not disprove these theories, it provides a plausible and empirically grounded model that does not require the presence of these special social structures. The other theories must, therefore, appeal to different sorts of data in order to support their claims. Cities are one of most exceptional and enduring of human inventions. Most great cities are exceptions in their own right: a New Yorker feels out of place in Los Angeles, Paris or Shanghai. However, this exceptionalism may be more due to our attention to human-scale details than to the underlying structures. In this paper, we have presented a generative theory that accounts for observed scaling in urban growth as a function of social-tie density and the diffusion of information across those ties. It is our hope that this provides both a foundation for the commonalities across all cities and a beginning point for which divergence between specific cities can be explored.

Methods

Data sets. All data sets used for analysis in this paper are publicly available. We collected data from the official websites of US Centre for Disease Control and Prevention, US Census Bureau and the Statistical Office of the European Union. The detailed information for each data set is provided in the paper.

Diffusion models. Assuming that the spread of information and disease are archetypes of simple contagions, for the simple exposure contagion case, we run the SI model[37,38] on networks generated by our model, and measure the speed at which the infection reaches a finite fraction of the population. We start by generating networks according to the process described in the introduction and then randomly pick 1% of the nodes as seeds (that is, initial infected nodes). The probability of an infection at a given time-step, to spread from an infected to a susceptible node, is denoted as ϵ, which we fix to be $\epsilon = 1 \times 10^{-2}$. The simulation terminates at the point when 10% of the population is in the infected state. The networks generated are snapshots at different densities ρ and as before we vary the size of the grid N.

For the complex contagion case, we adopt the complex contagion model[39]. We assume that 10% of the population follows a simple contagion process: an individual can be infected by a single infected neighbour; the remaining 90% of the population follows a complex contagion process: an individual is infected, if at least two of its neighbours are infected. The rest of the simulation is identical to the simulation with the SI model, and we measure the time steps required to infect 10% of the population.

Denoting $S(\rho)$ as the number of time steps taken to infect 10% of the population, the mean spreading rate $R(\rho)$ shown in Fig. 3 is computed using:

$$R(\rho) = \frac{\rho}{S(\rho)}. \tag{6}$$

Assuming that the mean spreading rate is proportional to the network density (that is, $R(\rho) \propto T(\rho)$), we also fit the data to the form

$$S(\rho) \sim \frac{\rho}{\frac{1}{k}T(\rho)} = \frac{k}{\ln\rho + C'}, \tag{7}$$

where k is a constant and $T(\rho)$ follows from equation (5). (Cf. Supplementary Note 6 and Supplementary Fig. S7).

References

1. Crane, P. & Kinzig, A. Nature in the metropolis. *Science* **308,** 1225 (2005).
2. US Census Bureau. Population, Housing Units, Area Measurements, and Density: 1790 to 2000, PHC-3-1 (2012).
3. Glaeser, E. L., Kolko, J. & Saiz, A. Consumer city. *J. Econ. Geog.* **1,** 27–50 (2001).
4. Smith, A. *The wealth of nations (1776)* (New York: Modern Library 740, 1937).
5. Milgram, S. The experience of living in cities. *Crowding and Behavior* **167,** 41 (1974).
6. Krugman, P. On the number and location of cities. *Eur. Econ. Rev.* **37,** 293–298 (1993).
7. Becker, G., Glaeser, E. & Murphy, K. Population and economic growth. *Am. Econ. Rev.* **89,** 145–149 (1999).
8. Fujita, M., Krugman, P. & Venables, A. *The Spatial Economy* (MIT Press, 1999).
9. Bettencourt, L., Lobo, J., Helbing, D., Kuhnert, C. & West, G. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl Acad. Sci.* **104,** 7301 (2007).
10. Bettencourt, L. & West, G. A unified theory of urban living. *Nature* **467,** 912–913 (2010).
11. Jaffe, A., Trajtenberg, M. & Henderson, R. Geographic localization of knowledge spillovers as evidenced by patent citations. *Q. J. Econ.* **108,** 577 (1993).
12. Audretsch, D. & Feldman, M. R&D spillovers and the geography of innovation and production. *Am. Econ. Rev.* **86,** 630–640 (1996).
13. Anselin, L., Varga, A. & Acs, Z. Local geographic spillovers between university research and high technology innovations. *J. Urban. Econ.* **42,** 422–448 (1997).
14. Arbesman, S., Kleinberg, J. & Strogatz, S. Superlinear scaling for innovation in cities. *Phys. Rev. E* **79,** 16115 (2009).
15. Leskovec, J., Lang, K., Dasgupta, A. & Mahoney, M. Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* **6,** 29–123 (2009).
16. Ahn, Y., Bagrow, J. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466,** 761–764 (2010).
17. Mucha, P., Richardson, T., Macon, K., Porter, M. & Onnela, J. Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328,** 876 (2010).
18. Expert, P., Evans, T., Blondel, V. & Lambiotte, R. Uncovering space-independent communities in spatial networks. *Proc. Natl Acad. Sci.* **108,** 7663 (2011).
19. Onnela, J., Arbesman, S., Gonzalez, M., Barabasi, A. & Christakis, N. Geographic constraints on social network groups. *PLoS ONE* **6,** e16939 (2011).
20. Brummitt, C., Gomez-Lievano, A., Goudemand, N. & Haslam, G. Hunting for keys to innovation: the diversity and mixing of occupations do not explain a citys patent and economic productivity. *Santa Fe Institute Technical Report* (2012).
21. Granovetter, M. The strength of weak ties. *Am. J. Sociol.* **78,** 1360–1380 (1973).
22. Granovetter, M. The impact of social structure on economic outcomes. *J. Econ. Persp.* **19,** 33–50 (2005).
23. Burt, R. *Structural holes: The social structure of competition* (Harvard University Press, 1995).
24. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328,** 1029 (2010).
25. Allen, T. *Managing The Flow Of Technology: Technology Transfer And The Dissemina-tion Of Technological Information Within The R&D Organization* (MIT Press, 2003).
26. Reagans, R. & Zuckerman, E. Networks, diversity, and productivity: the social capital of corporate R&D teams. *Organization Sci.* **12,** 502–517 (2001).
27. Wirth, L. Urbanism as a way of life. *Am. J. Sociol.* 1–24 (1938).
28. Hagerstrand, T. *The Propagation of Innovation Waves* (Royal University of Lund, 1952).
29. Hagerstrand, T. *Migration and Area* **13,** 27–158 (1957).
30. Florida, R. *The Rise of the Creative Class and How It's Transforming Work, Leisure, Community and Everyday Life* (Basic Books, 2004).
31. Florida, R. *Cities and the Creative Class* (Routledge, 2004).
32. Florida, R. *The Flight of the Creative Class: The New Global Competition for Talent* (Harper Business, 2007).
33. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. & Tomkins, A. Geographic routing in social networks. *Proc. Natl Acad. Sci.* **102,** 11623 (2005).
34. Krings, G., Calabrese, F., Ratti, C. & Blondel, V. A gravity model for inter-city telephone communication networks *J. Stat. Mech.* doi:10.1088/1742-5468/2009/07/L07003 (2009).
35. Calabrese, F. *et al.* The connected states of america: Quantifying social radii of inuence. *Proc. of IEEE International Conference on Social Computing* 223–230 (2011).
36. Shalizi, C. Scaling and hierachy in urban economies. *Arxiv preprint arXiv:1102.4101* (2011).
37. Kermack, W. & McKendrick, A. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115,** 700–721 (1927).
38. Anderson, R. & May, R. *Infectious diseases of humans: dynamics and control* (Oxford University Press, 1991).
39. Centola, D. & Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113,** 702–734 (2007).
40. US Center for Disease Control (2013); URL http://www.cdc.gov/hiv/topics/surveillance/index.htm.
41. Commission, E. *et al.* Regulation (ec) no 1059/2003 of the european parliament and of the council of 26 may 2003 on the establishment of a common classi_cation of territorial units for statistics (nuts). *Official J. European Union* **21,** 2003 (2003).
42. Baldassare, M. & Wilson, G. More trouble in paradise: urbanization and the decline in suburban quality-of-life ratings. *Urban Affairs Rev.* **30,** 690–708 (1995).
43. Eagle, N., Pentland, A. & Lazer, D. Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci.* **106,** 15274–15278 (2009).
44. Wu, L., Waber, B., Aral, S., Brynjolfsson, E. & Pentland, A. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task. *Proc. of Intl. Conference on Information Systems.* http://aisel.aisnet.org/icis2008/127 (2008).
45. Pentland, A. *Honest signals: How They Shape Our World* (MIT Press, 2008).

Acknowledgements

We thank L.M.A. Bettencourt and C.A. Hidalgo for their insightful comments and help on the manuscript. This research is supported by Army Research Laboratory under Cooperative Agreement Numbers W911NF-09-2-0053 and W911NF-11-1-0363, the National Science Foundation under grant 0905645, from DARPA/Lockheed Martin Guard Dog Programme under PO 4100149822.

Autho contributions

All authors contributed equally to this work. W.P., M.C. and A.P. designed the study. W.P. peformed research. G.G. analysed and described the analytic model. C.K. and G.G. contributed new reagents. C.K., G.G., W.P., M.C. and A.P. wrote the paper.

Additional Information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunication

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Pan, W. *et al.* Urban characteristics attributable to density-driven tie formation. *Nat. Commun.* 4:1961 doi: 10.1038/ncomms2961 (2013).

doi:10.1038/nature13462

Strong contributions of local background climate to urban heat islands

Lei Zhao[1,2], Xuhui Lee[1,2], Ronald B. Smith[3] & Keith Oleson[4]

The urban heat island (UHI), a common phenomenon in which surface temperatures are higher in urban areas than in surrounding rural areas, represents one of the most significant human-induced changes to Earth's surface climate[1,2]. Even though they are localized hotspots in the landscape, UHIs have a profound impact on the lives of urban residents, who comprise more than half of the world's population[3]. A barrier to UHI mitigation is the lack of quantitative attribution of the various contributions to UHI intensity[4] (expressed as the temperature difference between urban and rural areas, ΔT). A common perception is that reduction in evaporative cooling in urban land is the dominant driver of ΔT (ref. 5). Here we use a climate model to show that, for cities across North America, geographic variations in daytime ΔT are largely explained by variations in the efficiency with which urban and rural areas convect heat to the lower atmosphere. If urban areas are aerodynamically smoother than surrounding rural areas, urban heat dissipation is relatively less efficient and urban warming occurs (and vice versa). This convection effect depends on the local background climate, increasing daytime ΔT by 3.0 ± 0.3 kelvin (mean and standard error) in humid climates but decreasing ΔT by 1.5 ± 0.2 kelvin in dry climates. In the humid eastern United States, there is evidence of higher ΔT in drier years. These relationships imply that UHIs will exacerbate heatwave stress on human health in wet climates where high temperature effects are already compounded by high air humidity[6,7] and in drier years when positive temperature anomalies may be reinforced by a precipitation–temperature feedback[8]. Our results support albedo management as a viable means of reducing ΔT on large scales[9,10].

The conversion of natural land to urban land causes several notable perturbations to the Earth's surface energy balance. Reduction of evaporative cooling is generally thought to be the dominant factor contributing to UHI. Anthropogenic heat release is an added energy input to the energy balance and should increase the surface temperature. Energy input by solar radiation will also increase if albedo is reduced in the process of land conversion. Buildings and other artificial materials can store more radiation energy in the daytime than can natural vegetation and soil; release of the stored energy at night contributes to night-time

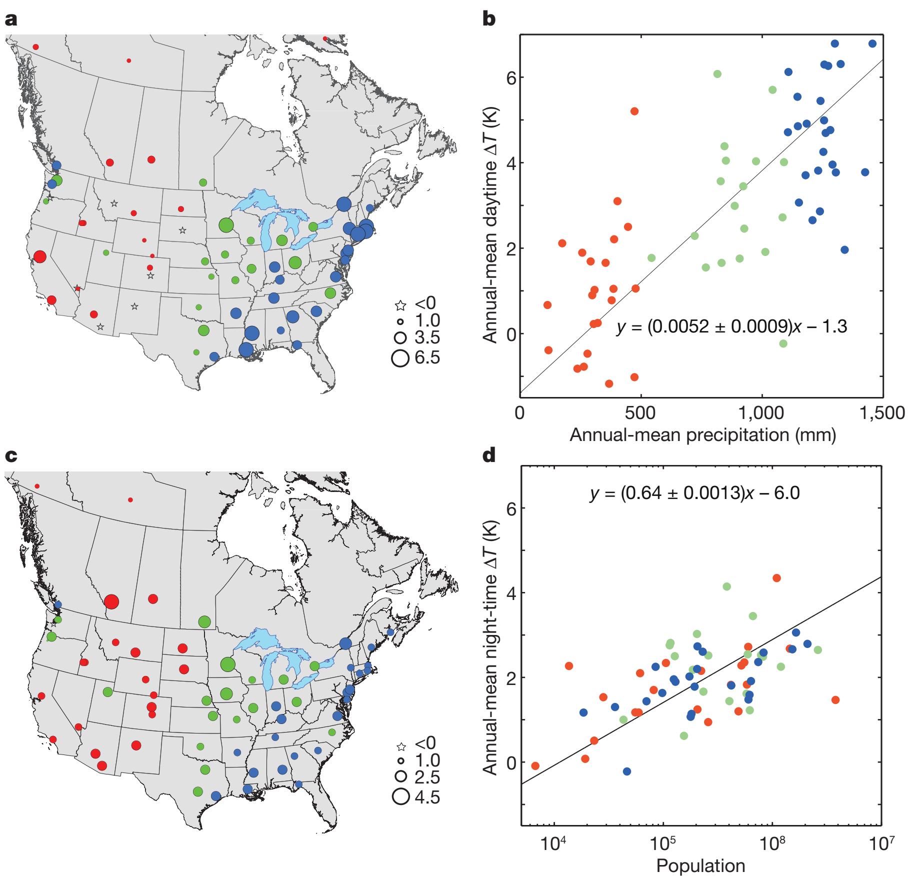


Figure 1 | Precipitation and population influences on MODIS-derived annual-mean UHI intensity. a, Map of daytime UHI (shown in K by symbol type/size). **b**, Dependence of daytime UHI on precipitation ($r = 0.74$, $P < 0.001$). **c**, Map of night-time UHI. **d**, Dependence of night-time UHI on population ($r = 0.54$, $P < 0.001$). Red, green and blue symbols denote cities with annual mean precipitations less than 500 mm, between 500 and 1,100 mm, and over 1,100 mm, respectively. Lines in **b** and **d** are linear regression fits to the data. Parameter bounds for the regression slope are the 95% confidence interval.

[1]Yale-NUIST Center on Atmospheric Environment, Nanjing University of Information Science and Technology, Nanjing 210044, China. [2]School of Forestry and Environmental Studies, Yale University, New Haven, Connecticut 06511, USA. [3]Department of Geology and Geophysics, Yale University, New Haven, Connecticut 06511, USA. [4]National Center for Atmospheric Research, Boulder, Colorado 80305, USA.

UHI. Finally, energy redistribution through convection between the surface and the atmospheric boundary layer can either increase or reduce ΔT, depending on whether the efficiency of convection over urban land is suppressed or enhanced relative to that over adjacent rural land. Although these concepts have been known for some time[11–13], a quantitative understanding of their roles across different climate regimes remains elusive[4].

The climatic context can be understood by posing the following question in a thought experiment: if two cities are built identically in terms of morphological and anthropogenic aspects but in different climates, will they have the same ΔT? The answer depends on time of day according to observations of surface temperature by the NASA MODIS satellite. For 65 selected cities in North America, the annual-mean midnight ΔT (surface temperature of urban core pixels minus that of rural pixels) is positively correlated with the logarithm of population (correlation coefficient, $r = 0.54$; confidence level, $P < 0.001$; Fig. 1d), but is invariant with climate, showing a statistically insignificant correlation with precipitation ($r = 0.05$, $P = 0.70$; Extended Data Fig. 1), solar radiation ($r = 0.15$, $P > 0.20$) and air temperature ($r = 0.20$, $P > 0.10$). However, the annual-mean midday ΔT is strongly correlated with precipitation ($r = 0.74$, $P < 0.001$; Fig. 1b) and has a weaker statistical dependence on population size than does the night-time ΔT ($r = 0.27$, $P = 0.027$; Extended Data Fig. 1). The night-time ΔT shows little spatial coherence (Fig. 1c), but the daytime ΔT has a discernible spatial pattern that follows precipitation gradients across the continent (Fig. 1a). Twenty-four of the cities are located in the humid southeast United States, which coincides roughly with the Köppen–Geiger temperate climate zone (Fig. 2a). Their daytime annual-mean ΔT is on average 3.9 K and is 3.3 K higher than that of the 15 cities in the dry region (Fig. 2d, e). By comparison, the night-time ΔT differs by 0.1 K between the two groups ($P > 0.60$; Fig. 2f, g). These results are in broad agreement with previous remote-sensing studies on UHI across biophysical and developmental gradients[14–17].

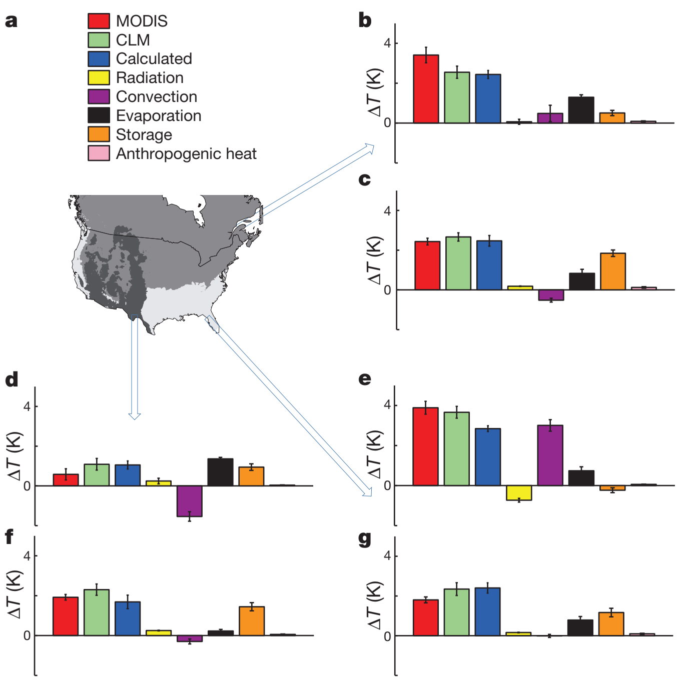


Figure 2 | Attribution of UHI intensity in three Köppen–Geiger climate zones. **a**, Map of climate zones: white, mild temperate/mesothermal climate; grey, continental /microthermal climate; dark grey, dry climate. **b**, **d**, **e**, Daytime values of MODIS and modelled ΔT and its component contributions in each of the three zones (see arrows). **c**, **f**, **g**, Night-time values in each of the three zones (see arrows). Green bars denote model-predicted ΔT and blue bars denote UHI intensity calculated as the sum of the component contributions. Error bars, 1 s.e. for each climate zone.

At first glance, the relationship with precipitation (Fig. 1b) seems consistent with the hypothesis that reduction in evaporative cooling in urban land is the main driver of daytime ΔT, because the denser vegetation in wet climate regions has a higher evaporation rate than the vegetation in dry climates. However, our model-based analysis does not support such an interpretation. In the model domain, ΔT is a perturbation signal to the surface temperature caused by biophysical contrast between rural and urban land units in the same model grid cell[18]. This signal is further decomposed, using the method described in ref. 19, into contributions from changes in radiation balance, evaporation, convection efficiency and heat storage, and from anthropogenic heat addition (Fig. 2). The credibility of the model is supported by the reasonable agreement of the modelled ΔT with the MODIS ΔT ($r = 0.31$, $P < 0.02$ for daytime; $r = 0.30$, $P < 0.025$ for night time) and by its accurate depiction of the relationship between night-time ΔT and albedo (Extended Data Fig. 4). Furthermore, the model has reproduced the observed positive correlation between the daytime ΔT and precipitation (Fig. 3a).

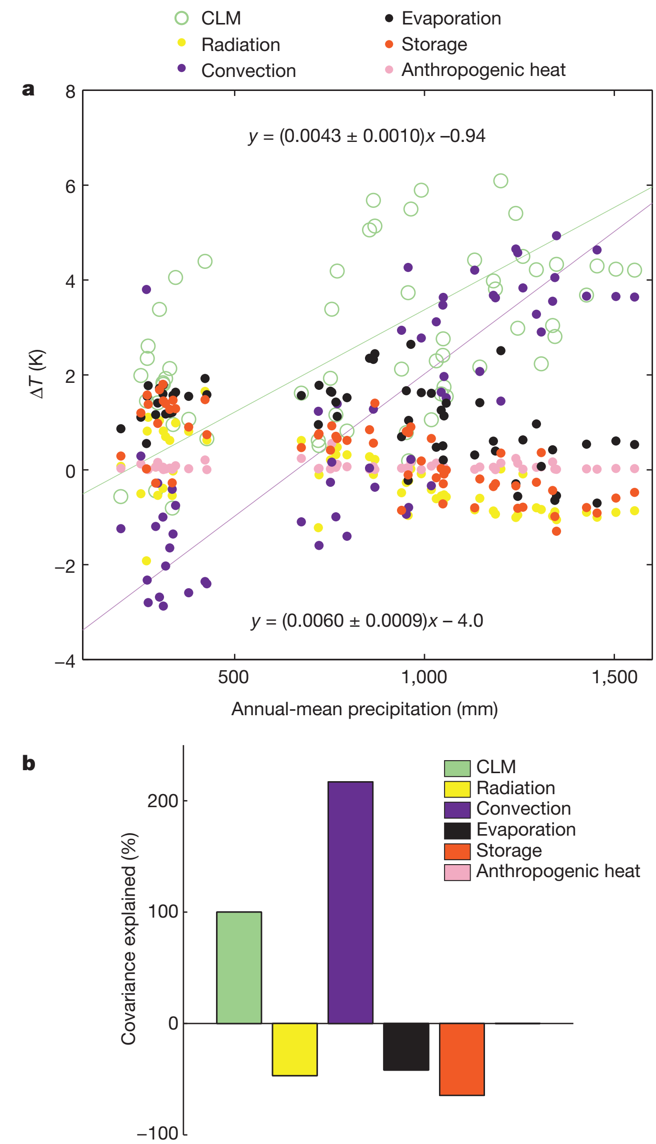


Figure 3 | Relationship between model-predicted daytime ΔT and precipitation among the cities. **a**, Correlation of ΔT and the individual biophysical components with annual-mean precipitation. Lines are linear regression fits to the corresponding data. Parameter bounds for the regression slope are the 95% confidence interval. **b**, ΔT–precipitation covariance explained by different biophysical factors. Note that the covariance explained by the anthropogenic heat term is negligibly small.

We find that it is the changes in convection efficiency (associated with aerodynamic resistance changes), rather than those in evapotranspiration, that control the daytime ΔT–precipitation spatial covariance among the cities (Fig. 3). In the humid climate (the Köppen–Geiger temperate climate zone), convection is less efficient at dissipating heat from urban land than from rural land, and the associated temperature increase is 3.0 ± 0.3 K, which dominates the overall ΔT (Fig. 2e). At these locations, the rural land is in general densely vegetated, owing to ample precipitation, and is aerodynamically rough. Quantitatively, this difference is manifested in a lower aerodynamic resistance to sensible heat diffusion in the rural land ($39\,\mathrm{s\,m^{-1}}$) than in the urban land ($62\,\mathrm{s\,m^{-1}}$). Measured in terms of aerodynamic resistance, urbanization has reduced the convection efficiency by 58%.

The opposite occurs in the dry climate zone, where urban land is rougher than rural land and has enhanced convection efficiency. The result is actually a cooling effect (Fig. 2d). In this zone, the urban landscape has lower aerodynamic resistance ($53\,\mathrm{s\,m^{-1}}$) than the adjacent rural land ($66\,\mathrm{s\,m^{-1}}$), which is typically inhabited by vegetation of low stature such as shrubs, sagebrushes and grasses. On average, the urban land is about 20% more efficient in removing heat from the surface by convection than is the rural land. The average cooling signal is -1.5 ± 0.2 K. In a few of the cities, convection is much more efficient than in the surrounding natural land, such that ΔT becomes negative (Figs 1a and 3a). It has been suggested that negative ΔT, a phenomenon known as 'urban heat sink', arises from evaporative cooling of trees and lawns planted in the city[15–17]. Our explanation seems more logical, because the MODIS urban temperature comes from the urban core pixels with negligible amounts of vegetation cover (enhanced vegetation index, <0.18) and the urban land unit in the climate model is completely free of vegetation. An analogous situation exists in a semi-arid plantation forest where trees serve as efficient 'heat convectors', leading to a lower surface temperature than in the adjacent smoother shrub land[20].

At night, release of the stored heat is the dominant contributor to ΔT across all three climate zones (Fig. 2c, f and g). The dependence on population size (Fig. 1d), which is an indicator of the city's horizontal dimension, can be understood in light of these results. At night, the released heat is trapped in a very shallow atmospheric boundary layer. As air moves across the urban land, it will accumulate more heat with increasing travel distance. Having a longer upwind fetch, that is, a longer distance between the upwind edge of the city and the point of observation, the centre of a larger city should experience stronger warming[11].

There is some evidence of precipitation control on interannual variability in the daytime ΔT for individual cities. For each city, we have calculated the linear regression slope of the annual daytime ΔT against the annual precipitation, and we refer to it as the temporal sensitivity to precipitation. Both the MODIS and the model data show a negative dependence of the sensitivity on site mean precipitation (Fig. 4b, d). Twenty-four cities have annual-mean precipitation exceeding 1,100 mm. According to the model, all of these cities, which are mostly distributed in the eastern United States, have negative temporal sensitivity (Fig. 4c), meaning higher ΔT in drier years. The mean temporal sensitivity of this group of cities is $-0.0021\,\mathrm{K\,mm^{-1}}$. The MODIS results are less consistent because of shorter data records, showing negative sensitivity for 42% of them (Fig. 4a, b).

To gain further insight into the interannual variability, we have compared the daytime ΔT–precipitation correlations for Billings in Montana (annual-mean precipitation, 353 mm) and Richmond in Virginia (1,183 mm). We choose these two cities because they have nearly the same morphological and biophysical specifications (Extended Data Table 1) and therefore are essentially identical in the model world. The sensitivity to precipitation is positive at Billings and negative at Richmond (Extended Data Fig. 2). In contrast to the spatial variations across North America (Fig. 3b), the ΔT interannual variability shown here is driven primarily by changes in surface evaporation (Extended Data Fig. 3).

Our results can be interpreted in the context of heatwave climatology. A measure of heatwave intensity is the degree of deviation, in multiples of standard deviation (North American mean value, $\sigma \approx 0.6$ K) of summertime temperature from the climatological mean[21]. For example, the 2003 European heatwave[8] is a rare event measured at 5σ. These statistical considerations are based on regional background climatology. Being an additional anomaly on this background condition, UHI will aggravate heat stress on human health. In the southeast United States, where the heat stress is already amplified by high air humidity[7], the daytime ΔT is equivalent to 7σ (Fig. 2e). The situation may be further

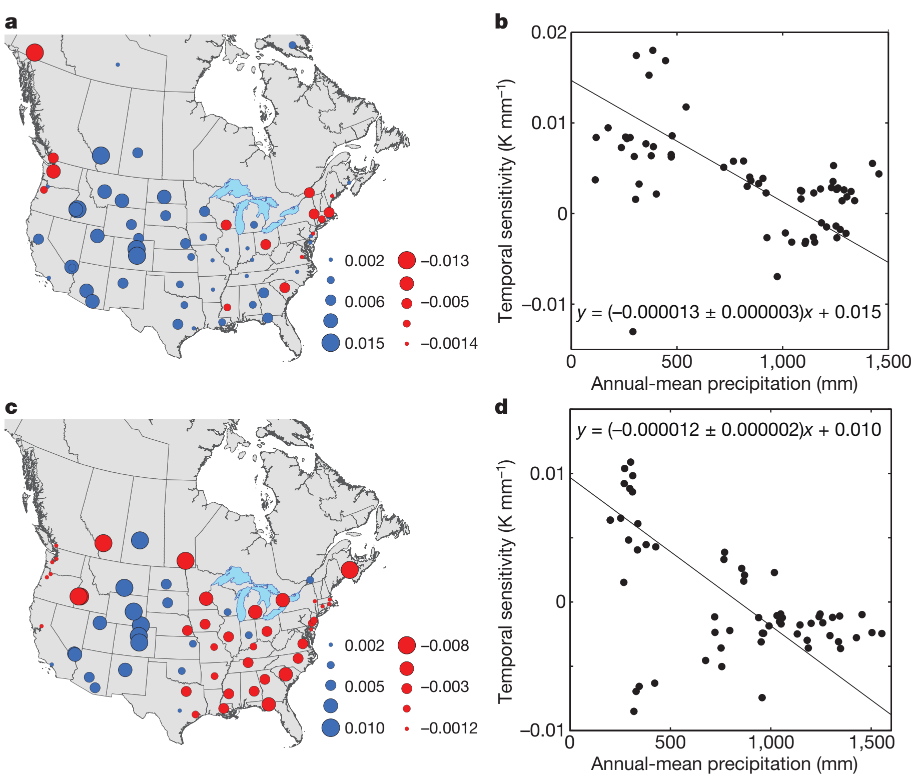


Figure 4 | Temporal sensitivity of UHI intensity to precipitation. **a**, **c**, Map of the temporal sensitivities (shown in $\mathrm{K\,mm^{-1}}$ by symbol size) according to MODIS (**a**) and the climate model (**c**). **b**, **d**, Dependence of MODIS (**b**) and model-predicted (**d**) temporal sensitivity on annual mean precipitation. The outlier city in the MODIS panels is Whitehorse in Yukon. The four outlier cities in the model panels are Boise and Nampa in Idaho, Winnipeg in Manitoba and Calgary in Alberta. Lines in **b** and **d** are linear regression fits to the data. Parameter bounds for the regression slope are the 95% confidence interval.

worsened in drier years when the positive temperature anomaly is likely to increase owing to a precipitation–temperature feedback[8]. Empirical evidence exists for such synergistic effects[22]. Using the temporal sensitivity of $-0.0021\ \mathrm{K\,mm^{-1}}$, a 500 mm reduction in the annual precipitation corresponds to an increase in the daytime ΔT by 1.1 K, or $\sim 2\sigma$. We caution that these numbers represent the upper bound of the UHI-added stress because UHI intensity at the screen height[23] (the height of air temperature observation at a standard weather station) and under all-sky conditions should be smaller than our ΔT, which is for clear skies and for the surface. However, summertime ΔT is generally larger than annual ΔT (refs 13–17, 24).

The health impact of heatwaves is one factor that motivates the growing efforts to mitigate UHI. According to our results, a strategy that focuses on reducing anthropogenic heat would bring virtually no relief, but this might be because of the primitive anthropogenic heat scheme in the model[18]. Managing the convection efficiency or heat storage of urban land does not seem viable, even though these are large contributors to ΔT, because it would require fundamental changes to the urban morphology, such as a city-wide increase in building height. However, efforts to increase urban albedo have the promise of producing measurable results on a large scale. For the cities in the southern United States, the reduction of net radiation loading amounts to a daytime cooling effect of 0.7 K (Fig. 2e). In the model, this reduction is caused by the fact that these cities have an average albedo that is 0.06 higher than the surrounding rural land. This albedo difference is modest, considering that phasing in reflective roofs in Chicago[25] has already increased the city-wide albedo by ~ 0.02 and that some cool-roof implementations[10] aim to increase the urban–rural albedo contrast by as much as 0.6. Albedo increases have little direct effect on the night-time UHI (Fig. 2g) but may have an indirect cooling benefit through the reduction in the daytime heat storage and, therefore, less heat release from storage at night[16,26,27]. The negative correlation between the night-time ΔT and urban–rural albedo contrast[16] (Extended Data Fig. 4) can be viewed as empirical evidence of this indirect benefit.

METHODS SUMMARY

MODIS data. We calculated the annual-mean ΔT using the MODIS-Aqua eight-day composite land surface temperature from 2003 to 2012. The night-time and daytime ΔT were determined at 1:30 and 13:30 local time, respectively. Data were collected at 65 cities distributed across the United States and Canada. For each city, we paired pixels in the centre of the city with those outside the city to determine ΔT.

Climate model. We used the Community Earth System Model[28] to simulate UHI. The model grid cell consists of urban and rural land units whose surface energy balance variables are calculated using a single-layer urban surface parameterization and a standard land surface scheme, respectively[18,29]. We ran the model for 33 yr of simulation time from 1972 to 2004 after a 60 yr spin-up. The forcing data are an atmospheric reanalysis product validated against various observations[30]. The simulation was conducted at the finest resolution supported by the model (0.23° longitude $\times$ 0.31° latitude) to resolve individual cities. The surface skin temperature was determined from the emitted long-wave radiation. To be consistent with the MODIS observations, we used the modelled data at 1:00 and 13:00 local time and under clear-sky conditions to compute the annual ΔT and to perform the surface energy balance analysis.

Attribution of UHI. Attribution of UHI is accomplished by a surface energy balance analysis. The total ΔT is partitioned, using the method of ref. 19, into contributions from the differences, between the urban and rural land units, in surface radiation balance, convection efficiency, evapotranspiration and heat storage, and from anthropogenic heat addition. The perturbation to the radiation balance results mainly from albedo contrast and also includes a minor part associated with surface emissivity change. The analysis was done separately for daytime and night time using the relevant forcing and prognostic model variables.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 January; accepted 7 May 2014.

1. Kalnay, E. & Cai, M. Impact of urbanization and land-use change on climate. *Nature* **423,** 528–531 (2003).
2. Zhou, L. M. *et al.* Evidence for a significant urbanization effect on climate in China. *Proc. Natl Acad. Sci. USA* **101,** 9540–9544 (2004).
3. Grimm, N. B. *et al.* Global change and the ecology of cities. *Science* **319,** 756–760 (2008).
4. Voogt, J. A. & Oke, T. R. Thermal remote sensing of urban climates. *Remote Sens. Environ.* **86,** 370–384 (2003).
5. Taha, H. Urban climates and heat islands: albedo, evapotranspiration, and anthropogenic heat. *Energy Build.* **25,** 99–103 (1997).
6. Fischer, E. M. & Schär, C. Consistent geographical patterns of changes in high-impact European heatwaves. *Nature Geosci.* **3,** 398–403 (2010).
7. Smith, T. T., Zaitchik, B. F. & Gohlke, J. M. Heat waves in the United States: definitions, patterns and trends. *Clim. Change* **118,** 811–825 (2013).
8. Schär, C. *et al.* The role of increasing temperature variability in European summer heatwaves. *Nature* **427,** 332–336 (2004).
9. Akbari, H., Menon, S. & Rosenfeld, A. Global cooling: increasing world-wide urban albedos to offset CO_2. *Clim. Change* **94,** 275–286 (2009).
10. Georgescu, M., Moustaoui, M., Mahalov, A. & Dudhia, J. Summer-time climate impacts of projected megapolitan expansion in Arizona. *Nature Clim. Change* **3,** 37–41 (2013).
11. Oke, T. R. The energetic basis of the urban heat-island. *Q. J. R. Meteorol. Soc.* **108,** 1–24 (1982).
12. Grimmond, S. Urbanization and global environmental change: local effects of urban warming. *Geogr. J.* **173,** 83–88 (2007).
13. Arnfield, A. J. Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island. *Int. J. Climatol.* **23,** 1–26 (2003).
14. Roth, M., Oke, T. R. & Emery, W. J. Satellite-derived urban heat islands from 3 coastal cities and the utilization of such data in urban climatology. *Int. J. Remote Sens.* **10,** 1699–1720 (1989).
15. Imhoff, M. L., Zhang, P., Wolfe, R. E. & Bounoua, L. Remote sensing of the urban heat island effect across biomes in the continental USA. *Remote Sens. Environ.* **114,** 504–513 (2010).
16. Peng, S. S. *et al.* Surface urban heat island across 419 global big cities. *Environ. Sci. Technol.* **46,** 696–703 (2012).
17. Clinton, N. & Gong, P. MODIS detected surface urban heat islands and sinks: global locations and controls. *Remote Sens. Environ.* **134,** 294–304 (2013).
18. Oleson, K. Contrasts between urban and rural climate in CCSM4 CMIP5 climate change scenarios. *J. Clim.* **25,** 1390–1412 (2012).
19. Lee, X. *et al.* Observed increase in local cooling effect of deforestation at higher latitudes. *Nature* **479,** 384–387 (2011).
20. Rotenberg, E. & Yakir, D. Contribution of semi-arid forests to the climate system. *Science* **327,** 451–454 (2010).
21. Hansen, J., Sato, M. & Ruedy, R. Perception of climate change. *Proc. Natl Acad. Sci. USA* **109,** E2415–E2423 (2012).
22. Li, D. & Bou-Zeid, E. Synergistic interactions between urban heat islands and heat waves: the impact in cities is larger than the sum of its parts. *J. Appl. Meteorol. Climatol.* **52,** 2051–2064 (2013).
23. Gallo, K. P., Adegoke, J. O., Owen, T. W. & Elvidge, C. D. Satellite-based detection of global urban heat-island temperature influence. *J Geophys Res.* **107,** 4776 (2002).
24. Tran, H., Uchihama, D., Ochi, S. & Yasuoka, Y. Assessment with satellite data of the urban heat island effects in Asian mega cities. *Int. J. Appl. Earth Obs. Geoinf.* **8,** 34–48 (2006).
25. Mackey, C. W., Lee, X. & Smith, R. B. Remotely sensing the cooling effects of city scale efforts to reduce urban heat island. *Build. Environ.* **49,** 348–358 (2012).
26. Oleson, K. W., Bonan, G. B., Feddema, J. & Vertenstein, M. An urban parameterization for a global climate model. Part II: sensitivity to input parameters and the simulated urban heat island in offline simulations. *J. Appl. Meteorol. Climatol.* **47,** 1061–1076 (2008).
27. Rosenzweig, C. *et al.* Mitigating New York City's heat island: integrating stakeholder perspectives and scientific evaluation. *Bull. Am. Meteorol. Soc.* **90,** 1297–1312 (2009).
28. Hurrell, J. W. *et al.* The Community Earth System Model: a framework for collaborative research. *Bull. Am. Meteorol. Soc.* **94,** 1339–1360 (2013).
29. Oleson, K. *et al. Technical Description of Version 4.0 of the Community Land Model (CLM)* 257. Report No. NCAR/TN-478+STR (NCAR, 2010).
30. Qian, T. T., Dai, A., Trenberth, K. E. & Oleson, K. W. Simulation of global land surface conditions from 1948 to 2004. Part I: forcing data and evaluations. *J. Hydrometeorol.* **7,** 953–975 (2006).

Acknowledgements This research was supported by the Ministry of Education of China (grant PCSIRT), the Yale Climate and Energy Institute, the Yale Institute of Biospheric Studies, and a Yale University Graduate Fellowship. K.O. acknowledges support from NASA grant NNX10AK79G (the SIMMER project) and the NCAR WCIASP. NCAR is sponsored by the US National Science Foundation. The model simulations were supported by the Yale University Faculty of Arts and Sciences High Performance Computing Center.

Author Contributions X.L. designed the research. L.Z. carried out the model simulation and data analysis. R.B.S. contributed ideas to the research design. K.O. contributed ideas to the model simulation. X.L. and L.Z. drafted the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.L. (xuhui.lee@yale.edu).

METHODS

MODIS LST, precipitation and population data. The UHI temperature difference ΔT can be defined using shelter-derived air temperature or a satellite-derived radiative surface temperature. The former suffers from inhomogeneity in the urban landscape. The latter is a valuable spatial average, but is influenced by the emissivity of the surface. Neither exactly matches the human experience of UHI as an individual walks across the rural or urban landscape. For the purpose of comparing different cities, the surface temperature approach is easier and more stable. Nichol *et al.* (ref. 31) showed that the correlation between the surface and screen-height air temperature can be weak on neighbourhood scales and improves considerably at the scale of urban–rural transition.

MODIS Aqua land surface temperature (LST) data obtained at 65 cities in the United States and Canada were used in this analysis. This is an eight-day clear-sky composite data set. The spatial resolution is 1 km. The satellite overpass times are approximately 13:30 and 1:30 local time, which are close to the times of daily maximum and minimum temperature, and the measurement therefore gives a better representation of the diurnal range of ΔT than does that of the other MODIS satellite, Terra. According to the product quality control flag, the data we used have an average LST error less than or equal to 2 K. While selecting urban–rural paired pixels, we avoided the rural pixels that have large elevation differences and large latitude differences relative to the urban core. Specifically, the upper thresholds for elevation difference and latitude difference are 100 m and 0.1°. Nine urban pixels (3×3) were selected in the city centre, paired with 1–3 patches of 9–49 pixels each (3×3 to 7×7) in the surrounding rural land. Because of topographic and latitudinal limitations, the number of rural pixels varied (one patch for 15 cities, two patches for 41 cities and three patches for 9 cities). The magnitude of ΔT is insensitive to the number of urban–rural pixels. Fixing the number of urban and rural pixels for all the cities to one 3×3 patch of pixels altered ΔT by at most 0.6 K. All the pixels selected were validated by the MODIS land cover map and cross-checked against Google Earth. Rural pixels are classified in the MODIS land cover map as natural surfaces such as forests, grassland, cropland and bare soils. To avoid high bias of UHI, we excluded water pixels. Urban pixels are classified in the MODIS land cover map as urban and built-up surface. The resulting ΔT represents the difference between the city core and minimally developed land outside the city. The annual mean values were calculated based on the 10 yr time series of the MODIS LST (2003–2012). Linear gap filling was done for short periods of missing values to minimize the impact of missing data on the annual means. If there are more than three consecutive missing values, we excluded that year.

Cities were chosen so that each state, province or territory was represented by at least one city, with the exception of four provinces and a territory in Canada (Nova Scotia, Prince Edward Island and Newfoundland and Labrador; Extended Data Table 2). The chosen cities are large enough to be resolved by the climate model, except for five small cities (Helena, Montana; Augusta, Maine; Whitehorse, Yukon; Yellowknife, Northwest Territories; Iqaluit, Nunavut). These cities span a population range of 7,000–379,300. In addition, we avoided the cities on hilly terrain.

The US precipitation data were obtained from PRISM (PRISM Climate Group, Oregon State University; http://prism.oregonstate.edu). The precipitation data for cities in Canada were obtained from Environment Canada (http://climate.weather.gc.ca/). The PRISM data sets are elevation-corrected grid estimates of monthly, yearly and event-based climatic variables. The precipitation data for Canadian cities are station measurements.

The population data were obtained from the US Census 2010 (http://quickfacts.census.gov) and Canada 2011 Census from Statistics Canada (http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/index-eng.cfm).

Climate model and simulation. We used NCAR's climate model CESM[28] (Community Earth System Model) to simulate the UHI in the United States and in Canada. In this model system, the land surface processes are represented by the Community Land Model[29] (CLM). We used CLM version 4.0. In CLM, the land surfaces are categorized into five land units: vegetation, glacier, wetland, urban and lake. Each grid cell can have one or more of these land units. The surface radiation and energy balance equations are solved separately for these land units, and the results are aggregated to yield grid cell means. Specifically, the urban land unit is modelled using a 'canyon' structure and consists of the following subsurfaces: roof, sunlit wall, shaded wall, and pervious (for example bare soil) and impervious (for example road, sidewalk and car park) canyon floor. It should be noted here that there is potentially more evaporation from the pervious canyon floor in the urban land unit than from the comparable bare soil in the rural land, because all of the water in the bare soil column is available for evaporation in the urban land unit. The vegetated land unit corresponds to nonurban or rural land. This land unit may contain up to 15 different plant functional types and bare soil.

The model was run in the offline mode (uncoupled from an active atmospheric model). The urban and rural parameterizations in each grid cell were driven by the same atmospheric forcing. The atmospheric forcing data used in this study is a careful reconstruction of the climatology from 1972 to 2004[30]. It was derived from a combination of the NCEP-NCAR reanalysis[32], observation-based analyses and observational records. Therefore, the data set has an improved accuracy compared with the NCEP-NCAR reanalysis. We ran the model for 33 yr from 1972 to 2004 after a 60 yr spin-up. The simulation was conducted at the finest resolution as a standard model configuration supported by this version of the model (0.23° longitude $\times$ 0.31° latitude), to resolve individual cities. We note that even at this finest resolution the grid cell is still large enough that the total urban area in a grid cell can be a combination of several urban areas. The surface skin temperature was determined from the emitted long-wave radiation for each land unit, with an emissivity of 0.88 for the urban land unit and 0.96 for the vegetated land unit. The urban emissivity is the mean value of the weighted averages of the emissivity values of the urban subsurfaces prescribed in the model for the selected cities. The rural emissivity is the mean value of the weighted average of the vegetation and soil emissivity.

To construct the UHI, urban and rural flux and state variables were extracted from the model output at the grid cells where the selected cities reside. CLM invokes the urban parameterization only if the urban area fraction exceeds a threshold of 0.1%. Therefore, five small cities used in the MODIS data analysis (Helena, Montana; Augusta, Maine; Whitehorse, Yukon; Yellowknife, Northwest Territories; Iqaluit, Nunavut) are neglected by the model.

We included only the modelled data at 1:00 and 13:00 local time each day in this analysis; these times were selected to match closely the MODIS overpass times. To replicate the MODIS clear-sky conditions, we excluded cloudy days whose clearness index[33] was less than 0.5. We then converted the daily values into eight-day averages. The gap filling and processes of calculating annual means are the same as for the MODIS data. Under all-sky conditions, the modelled UHI intensity is on average 0.59 K lower during the daytime and 0.02 K lower at night than the clear-sky values, and the pattern regarding the component contributions remains unchanged from the clear-sky plot (Fig. 2).

We note that the climate model cannot explicitly capture population dependence (Fig. 1). This is because population size is not a model parameter and the heat advection occurs at subgrid scales not resolved by its one-dimensional parameterization of land–atmosphere interactions.

Attribution of UHI. We used a surface energy balance analysis to isolate the contribution to the model-predicted ΔT from each individual biophysical factor associated with urban land conversion. In this analysis, the rural land is regarded as the base state, and urbanization is a perturbation to this base state. The perturbation signal is denoted by Δ. For example, $\Delta T = T_u - T_r$, where T_u is urban surface temperature and T_r is rural surface temperature within the same model grid cell. Following the method of ref. 19, the solution of the UHI intensity can be approximated by

$$\Delta T \approx \frac{\lambda_0}{1+f}\Delta R_n^* + \frac{-\lambda_0}{(1+f)^2}(R_n^* - Q_s + Q_{AH})\Delta f_1 + \frac{-\lambda_0}{(1+f)^2}(R_n^* - Q_s + Q_{AH})\Delta f_2 + \frac{-\lambda_0}{1+f}\Delta Q_s + \frac{\lambda_0}{1+f}\Delta Q_{AH} \qquad (1)$$

with

$$f = \frac{\lambda_0 \rho C_p}{r_a}\left(1 + \frac{1}{\beta}\right)$$

$$R_n^* = (1-a)K_\downarrow + L_\downarrow - (1-\varepsilon)L_\downarrow - \varepsilon\sigma T_a^4$$

$$\Delta f_1 = \frac{-\lambda_0 \rho C_p}{r_a}\left(1 + \frac{1}{\beta}\right)\frac{\Delta r_a}{r_a}$$

$$\Delta f_2 = \frac{-\lambda_0 \rho C_p}{r_a}\frac{\Delta\beta}{\beta^2}$$

where T is the surface temperature, $\lambda_0 = 1/4\varepsilon\sigma T^3$ is the local climate sensitivity, f is the energy redistribution factor, R_n^* is the apparent net radiation, ρ is the air density, C_p is the specific heat of air at constant pressure, r_a is the aerodynamic resistance to heat diffusion, β is the Bowen ratio, a is the surface albedo, $K_\downarrow$ is the incoming solar radiation, $L_\downarrow$ is the incoming long-wave radiation, ε is the surface emissivity, σ is the Stefan–Boltzmann constant, T_a is the air temperature at a reference height. In this analysis, we assume that r_a, β, R_n* and Q_s and Q_{AH} are parameters associated with

the external perturbation (land use conversion) and are independent of T; the partial derivative of these variables can then be calculated.

In equation (1), the terms on the right-hand side represent, in order from the first to the last, contributions from changes in radiation balance (term 1), aerodynamic resistance (term 2), Bowen ratio (term 3), and heat storage (term 4) and from anthropogenic heat addition (term 5). Because r_a is the resistance to sensible or convection heat flux, term 2 is essentially a measure of change in the convection efficiency between urban and rural land. In an abstract sense, changes in β (term 3) can result from changes in sensible heat flux (H), latent heat flux (LE) or both (see equation (2), below). In the present context of partial differentiation, however, H and LE are not independent because the delta term Δf_2 is evaluated with the net radiation R_n^* and other valuables held constant. Thus, a reduction in β is accomplished by channelling more radiation energy to the surface latent heat flux, and it is appropriate to attribute term 3 to changes in surface evaporative cooling.

The calculation was performed separately for 1:00 and 13:00 local time, with cloudy days omitted. Three sets of variables were used. The first set comes directly from the forcing data and includes precipitation, incoming solar radiation ($K_\downarrow$), reference-height air temperature (T_a; air temperature at 30 m above the surface), air pressure and downward long-wave radiation ($L_\downarrow$). The second set has model-predicted variables, including reflected short-wave radiation ($aK_\downarrow$), sensible heat flux (H), latent heat flux (LE), storage heat flux (Q_s) and anthropogenic heat flux (Q_{AH}). The third set of variables, including surface temperature (T), air density (ρ), Bowen ratio (β) and aerodynamic resistance (r_a) were derived from the forcing data and the model-predicted variables. Specifically, the Bowen ratio was calculated as

$$\beta = \frac{H}{\mathrm{LE}} \qquad (2)$$

and the aerodynamic resistance to heat diffusion was calculated from

$$r_a = \frac{\rho C_p (T - T_a)}{H} \qquad (3)$$

The aerodynamic resistance determined from equation (3) is the sum of the diffusion resistance in the atmospheric surface layer and the excess resistance associated with the thermal roughness[34,35]. The urban and rural land units within each model grid cell have the same forcing variables and have different values for the second and third sets of the variables. It should be noted here that the model underestimates the anthropogenic heat flux (Q_{AH}) owing to the primitive anthropogenic heat scheme. The total anthropogenic heat in the model includes only heating and air conditioning (HAC) fluxes, waste heat generated by HAC and the heat removed by air conditioning. These fluxes are based on some prescribed parameters in the surface data set of CLM and calculated heat transfer into and out of roofs and walls. The heat flux due to traffic is neglected by the current version of the model[36].

The sum of the component contributions is slightly lower than the modelled ΔT (Fig. 2) because high-order terms are ignored in the linearization of the surface long-wave radiation term of the energy balance equation and nonlinear interactions among the factors are omitted in the analysis. Comparison between model-predicted ΔT and calculated ΔT (sum of the individual contributions) reveals excellent correlation for daytime ($r = 0.88$, $P < 0.001$) and night time ($r = 0.55$, $P < 0.001$).

Covariance analysis. The covariance analysis was performed on modelled ΔT and its components against precipitation. Let C_R, C_H, C_{LE}, C_s and C_{AH} be the contributions from radiation, convection efficiency, evaporation, storage and anthropogenic heat, respectively (terms 1 to 5 in equation (1)). Equation (1) can be rewritten as

$$\Delta T = C_R + C_H + C_{LE} + C_s + C_{AH} + e$$

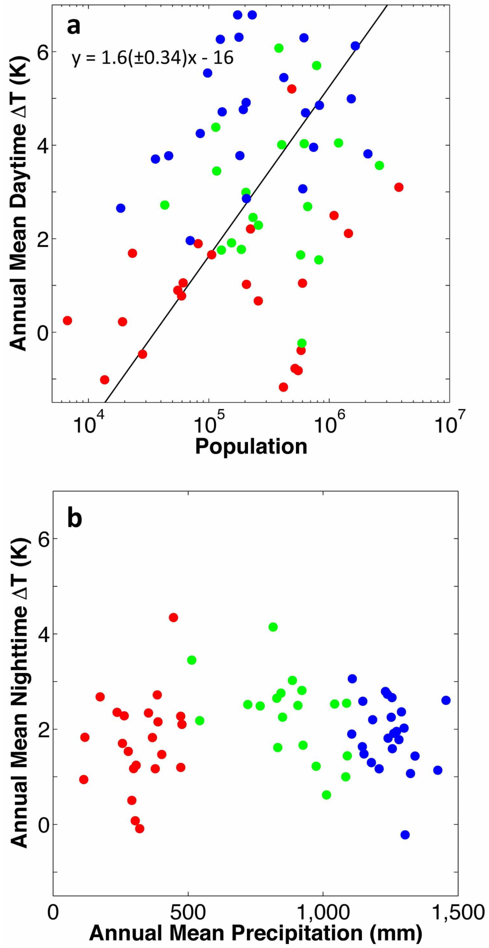
where e is an error term arising from nonlinear interactions. Because the covariance operation is linear, the ΔT–precipitation covariance is equal to the sum of the covariance between each component and precipitation

$$\begin{aligned}\mathrm{Cov}(\Delta T, P) = {} & \mathrm{Cov}(C_R, P) + \mathrm{Cov}(C_H, P) \\ & + \mathrm{Cov}(C_{LE}, P) + \mathrm{Cov}(C_s, P) \\ & + \mathrm{Cov}(C_{AH}, P) + \mathrm{Cov}(e, P)\end{aligned} \qquad (4)$$

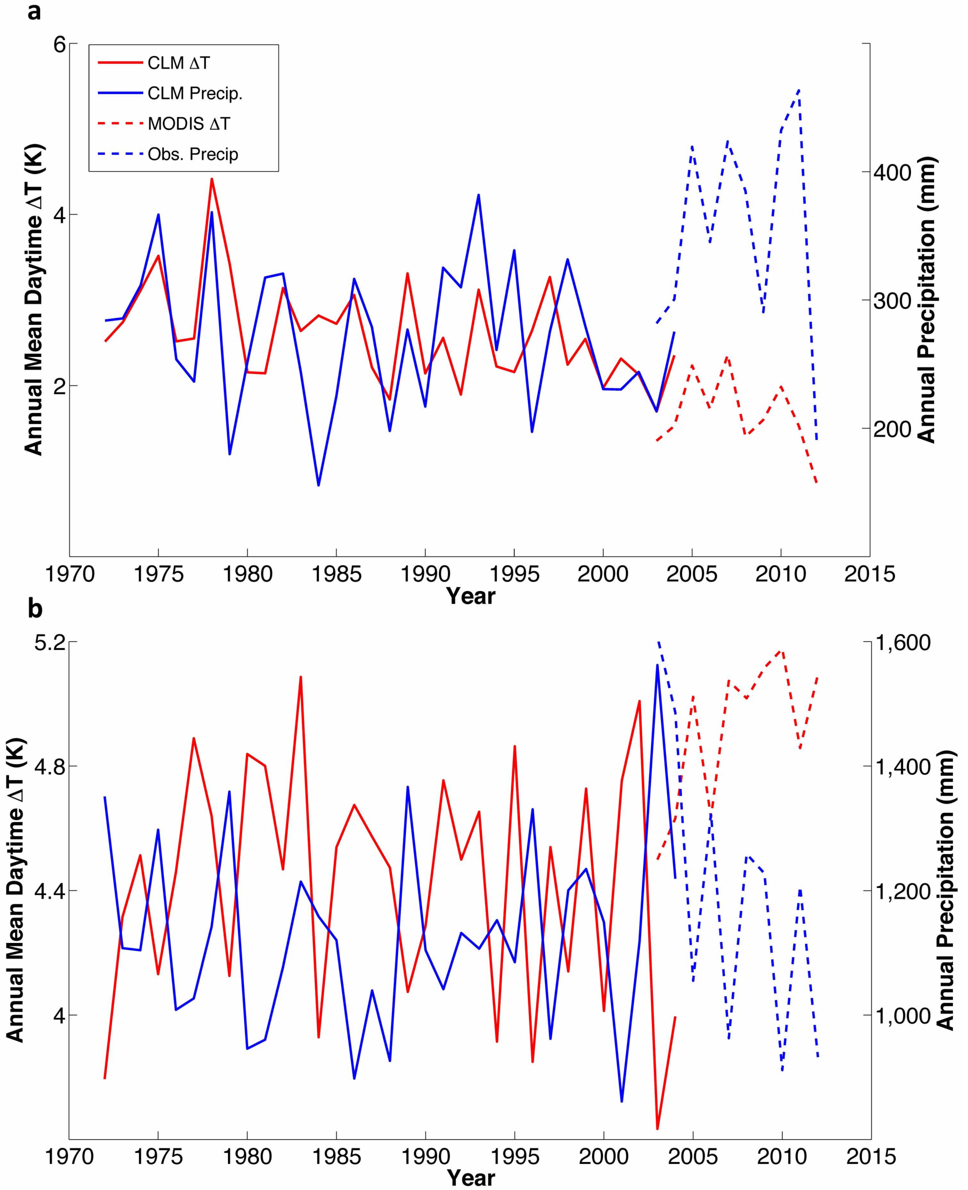
where P is precipitation. Equation (4) decomposes the total covariance between ΔT and precipitation into the covariance contribution from its five components and a residual error term. We presented covariance here rather than correlation coefficient because the correlation is not a linear operation. In Fig. 3 and Extended Data Fig. 3, we normalized the covariance between each component and precipitation by the total ΔT–precipitation covariance.

We applied this technique to the analysis of both spatial covariance and temporal covariance. In the analysis of spatial covariance, each data point is the climatic annual mean value of a city (Fig. 3). In the analysis of the temporal covariance at a city, each data point is the mean value for an individual year of that city (Extended Data Fig. 3).

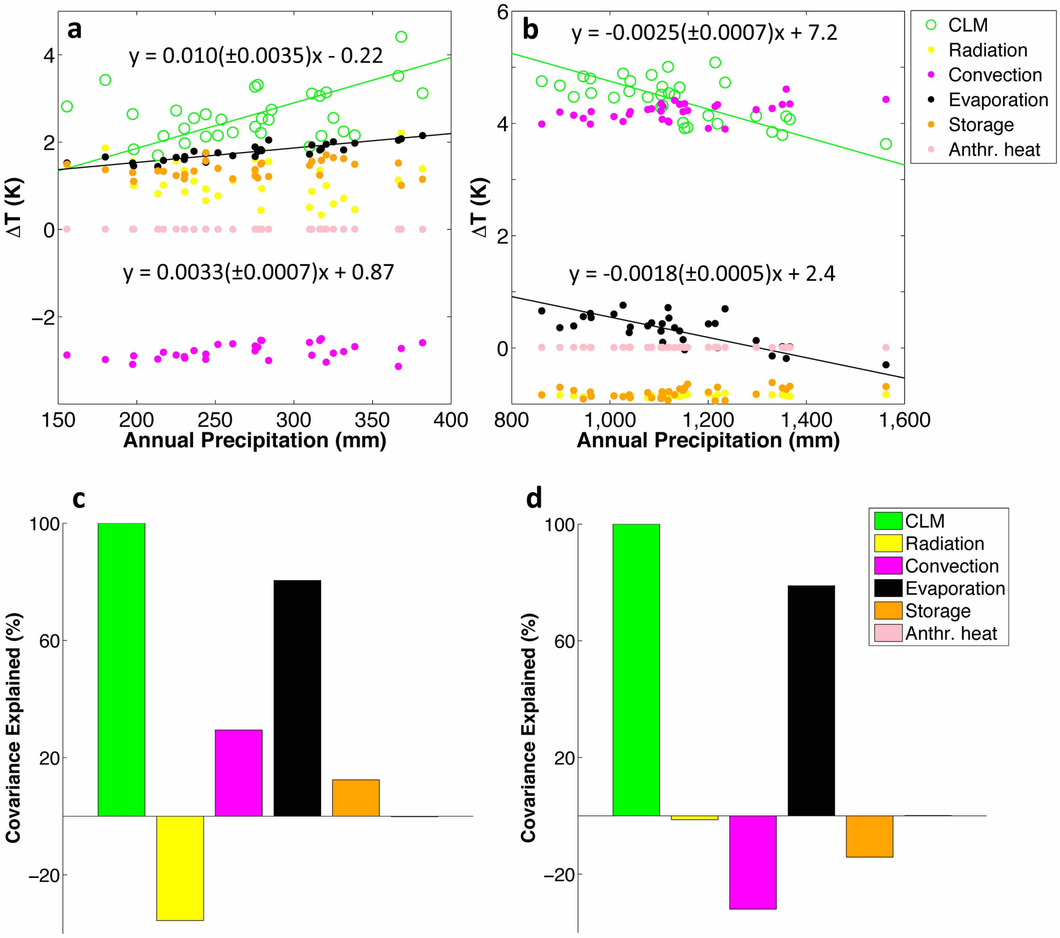
31. Nichol, J. *et al.* Urban heat island diagnosis using ASTER satellite images and 'in situ' air temperature. *Atmos. Res.* **94,** 276–284 (2009).
32. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–471 (1996).
33. Gu, L. H., Fuentes, J. D., Shugart, H. H., Staebler, R. M. & Black, T. A. Responses of net ecosystem exchanges of carbon dioxide to changes in cloudiness: results from two North American deciduous forests. *J. Geophys. Res., D, Atmospheres* **104,** 31421–31434 (1999).
34. Garratt, J. R. *The Atmospheric Boundary Layer* (Cambridge Univ. Press, 1994).
35. Voogt, J. A. & Grimmond, C. S. B. Modeling surface sensible heat flux using surface radiative temperatures in a simple urban area. *J. Appl. Meteorol.* **39,** 1679–1699 (2000).
36. Oleson, K. W., Bonan, G. B., Feddema, J. & Jackson, T. An examination of urban heat island characteristics in a global climate model. *Int. J. Climatol.* **31,** 1848–1865 (2011).



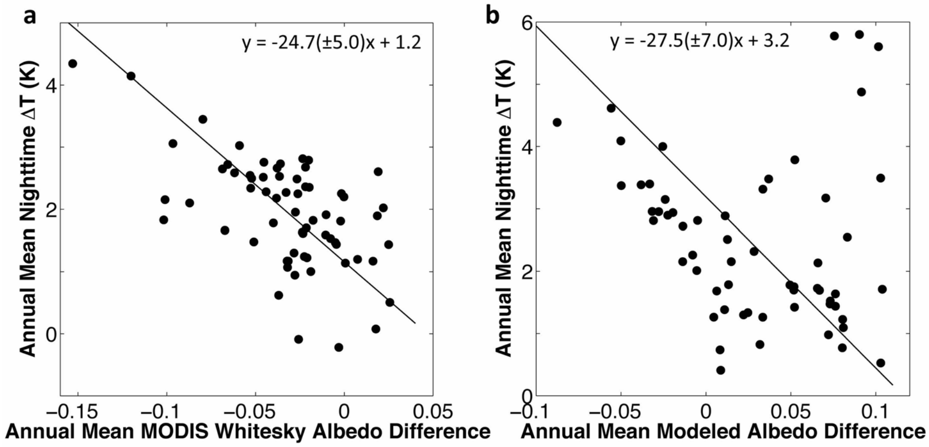
Extended Data Figure 1 | Precipitation and population influences on MODIS-derived annual mean UHI intensity. **a**, Dependence of daytime UHI on population size ($r = 0.27$, $P = 0.027$). **b**, Dependence of night-time UHI on precipitation ($r = 0.05$, $P = 0.70$). Red, green and blue symbols denote cities with annual mean precipitations less than 500 mm, between 500 and 1,100 mm, and over 1,100 mm, respectively. The solid line in **a** is the linear regression fit to the data. Parameter bounds for the regression slope are the 95% confidence interval.



Extended Data Figure 2 | Time series of MODIS and model-predicted daytime ΔT and annual precipitation. **a**, Billings, Montana. **b**, Richmond, Virginia.



Extended Data Figure 3 | Relationship between interannual variations in model-predicted daytime ΔT and precipitation. **a**, Correlation of ΔT and the individual biophysical components with annual precipitation at Billings, Montana. **b**, Same as in **a** except for Richmond, Virginia. **c**, ΔT–precipitation temporal covariance explained by different biophysical factors at Billings, Montana. **d**, Same as in **c** except for Richmond, Virginia. Lines are best linear regression fits to the data points. Parameter bounds for the regression slope are the 95% confidence interval.



Extended Data Figure 4 | Albedo influence on annual mean night-time UHI intensity. **a**, Dependence of night-time MODIS-derived UHI on white-sky albedo difference (that is, urban albedo minus rural albedo; $r = -0.60$, $P < 0.001$). **b**, Dependence of night-time modelled UHI on modelled albedo difference ($r = -0.56$, $P < 0.001$ excluding four outliers; $r = -0.18$, $P = 0.16$ with all data points). The four outliers in the upper right corner of **b** are coastal cities (Olympia, Washington; Seattle, Washington; Salem, Oregon; Vancouver, British Columbia) that have high biases of the modelled ΔT compared to the MODIS ΔT. Lines are linear regression fits to the data. Parameter bounds for the regression slope are the 95% confidence interval.

Extended Data Table 1 | Urban parameters of a city pair in CLM

City	Richmond	Billings
State	Virginia	Montana
Latitude (^{o})	37.53	45.79
Longitude (^{o})	-77.42	-108.54
Canyon Height/Width	0.48	0.48
Mean building height (m)	12	12
Roof thickness (m)	0.15	0.15
Wall thickness (m)	0.28	0.28
Wind height in canyon (m)	6	6
Roof fraction	0.55	0.50
Pervious road fraction	0.66	0.64
Emissivity (Impervious road)	0.91	0.91
Emissivity (pervious road)	0.95	0.95
Emissivity (roof)	0.65	0.65
Emissivity (wall)	0.91	0.91
Albedo (Impervious road)	0.13	0.13
Albedo (pervious road)	0.08	0.08
Albedo (roof)	0.30	0.30
Albedo (wall)	0.34	0.34
Roof thermal conductivity (W m^{-1} K^{-1})	0.84	0.84
Wall thermal conductivity (W m^{-1} K^{-1})	1.06	1.06
Impervious road thermal conductivity (W m^{-1} K^{-1})	1.67	1.67
Layers of impervious road	2	2
Roof heat capacity (MJ m^{-3} K^{-1})	0.76	0.76
Wall heat capacity (MJ m^{-3} K^{-1})	0.81	0.81
Impervious road heat capacity (MJ m^{-3} K^{-1})	2.06	2.06

Extended Data Table 2 | Size statistics for selected cities in the United States and in Canada

City, State/Province	Population	City size in CLM (km^2)	City, State/Province	Population	City size in CLM (km^2)
Albany, NY	9.79E+04	15.45	Louisville, KY	7.41E+05	45.75
Albuquerque, NM	5.53E+05	25.24	Madison, WI	2.33E+05	22.73
Atlanta, GA	4.20E+05	117.03	Minneapolis, MN	3.83E+05	90.58
Augusta, ME	1.86E+04	NA	Montgomery, AL	2.06E+05	12.99
Austin, TX	8.21E+05	63.16	Montreal, QC	1.65E+06	41.10
Baton Rouge, LA	2.30E+05	17.70	Nampa, ID	8.16E+04	6.56
Billings, MT	1.06E+05	2.14	Nashville, TN	6.36E+05	30.35
Bismarck, ND	6.13E+04	6.05	Oklahoma City, OK	5.80E+05	10.72
Boise, ID	2.06E+05	12.84	Olympia, WA	4.65E+04	12.94
Boston, MA	6.18E+05	86.42	Philadelphia, PA	1.53E+06	71.51
Calgary, AB	1.10E+06	21.48	Phoenix, AZ	1.45E+06	53.91
Casper, WY	5.53E+04	3.33	Pierre, SD	1.36E+04	0.89
Cheyenne, WY	5.95E+04	2.46	Portland, OR	5.94E+05	54.50
Colorado Springs, CO	4.16E+05	29.43	Providence, RI	1.78E+05	16.06
Columbia, SC	1.29E+05	2.56	Raleigh, NC	4.04E+05	37.60
Columbus, OH	7.87E+05	14.55	Richmond, VA	2.04E+05	8.00
Dallas, TX	1.20E+06	153.98	Sacramento, CA	4.89E+05	36.27
Denver, CO	6.00E+05	75.02	Saint John, NB	7.01E+04	1.42
Des Moines, IA	2.03E+05	32.56	Salem, OR	1.55E+05	14.41
Dover, DE	3.60E+04	1.84	Salt Lake City, UT	1.86E+05	57.87
Hartford, CT	1.25E+05	25.03	Saskatoon, SK	2.22E+05	13.07
Helena, MT	2.82E+04	NA	Seattle, WA	6.21E+05	108.04
Henderson, NV	2.58E+05	31.46	Springfield, IL	1.16E+05	12.97
Houston, TX	2.10E+06	151.09	Tallahassee, FL	1.81E+05	7.58
Indianapolis, IN	8.30E+05	37.56	Topeka, KS	1.27E+05	5.16
Iqaluit, NU	6.70E+03	NA	Toronto, ON	2.62E+06	215.29
Jackson, MS	1.74E+05	15.76	Trenton, NJ	8.49E+04	36.53
Jefferson City, MO	4.31E+04	3.54	Tucson, AZ	5.20E+05	27.73
Lansing, MI	1.14E+05	6.47	Vancouver, BC	6.04E+05	129.04
Las Vegas, NV	5.84E+05	31.46	whitehorse, YT	2.33E+04	NA
Lincoln, NE	2.58E+05	17.24	Winnipeg, MB	6.64E+05	47.08
Little Rock, AR	1.94E+05	13.64	Yellowknife, NT	1.92E+04	NA
Los Angeles, CA	3.79E+06	213.56			

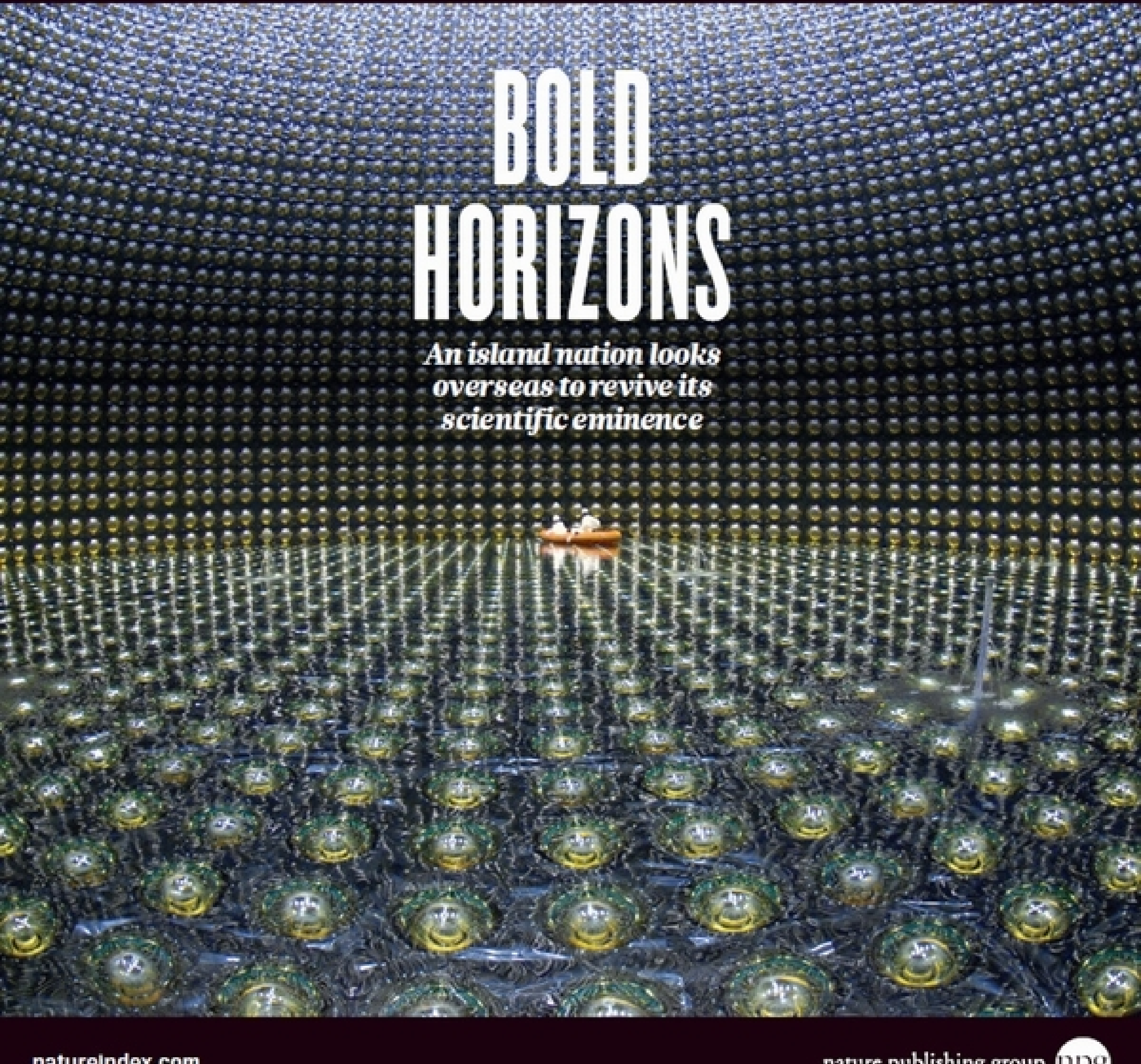
nature INDEX 2016 JAPAN

Is Japan playing catch-up in the collaboration game?

Inside the labs of the country's top 10 institutions

The rise and rise of trusted stalwarts and new niche performers

The tables



BOLD HORIZONS

An island nation looks overseas to revive its scientific eminence

nature publishing group npg

nature INDEX 2016 JAPAN

NATURE, VOL. 531, ISSUE NO. 7594 (MARCH 17, 2016)

The Japanese government recently presented a confronting picture of the state of science, technology and innovation in the country.

In a 2015 report on its Fifth Science and Technology Basic Plan, Japan's Council for Science, Technology and Innovation conceded its world standing in science and technology was falling. Government R&D investment growth has also stagnated in the past 10 years compared to spending by the world's other leading nations. "Our research papers are dropping in international rank, in quantitative and qualitative terms," the report said. "There have been delays in building an international network, and our science and technology activity is regrettably starting to fall behind the world leaders."

The Nature Index is a powerful tool to examine the nuances of this situation. While the database provides a broad overview of the country's research performance in the natural sciences, it also offers insights into the high-quality output and collaboration trends of individual institutions.

Japan is still a very strong scientific country with seven universities in the top 100 in the Nature Index. By some measures, the index reflects the government's recent assessment of the diminished state of science and innovation, it also reveals that some of Japan's institutions have increased their output of high-quality science.

Japan's overall output in the index has decreased by 12% since 2012, and the country is ranked in the database's top five countries. There is also evidence that the government's agenda to globalize universities and encourage international partnerships is working. In the past year, connections with nine of its top 10 collaborating countries have increased.

In our first index supplement dedicated to Japan, we present the performance of individual institutions in three categories.

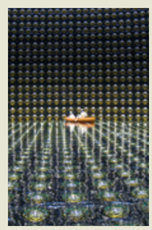
On page S114 we profile the country's top universities by weighted fractional count. These institutions are publishing the largest portion of Japan's best research. We look at the country's rising stars on page S118. These are the institutions with the biggest growth in high-quality output between 2012 and 2015. Some are small or new, but their improvement surpasses many older and better-resourced universities.

Japan's institutions have improved their collaborative reach and the article on page S127 looks at some institutions whose research partnerships have resulted in the most publications in high-impact journals, measured by a metric known as collaboration score. For more information on how Nature Index metrics are calculated, see page S136.

Nicky Phillips
Editor, Nature Index

CONTENTS

COVER IMAGE
The world's largest underground neutrino detector, Super-Kamiokande is hosted by the Institute of Cosmic Ray Research, part of the University of Tokyo.



KAMIOKA OBSERVATORY/ICRR/ THE UNIVERSITY OF TOKYO

NATURE INDEX 2016 JAPAN
The Nature Index 2016 Japan, a supplement to *Nature*, is produced by Nature Publishing Group, a division of Macmillan Publishers Ltd. This publication is based on data from the Nature Index, a website maintained by Nature Publishing Group and made freely available at natureindex.com.

Nature Editorial Offices
The Macmillan Building
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

CUSTOMER SERVICES
To advertise with the Nature Index, please visit natureindex.com/support
feedback@nature.com

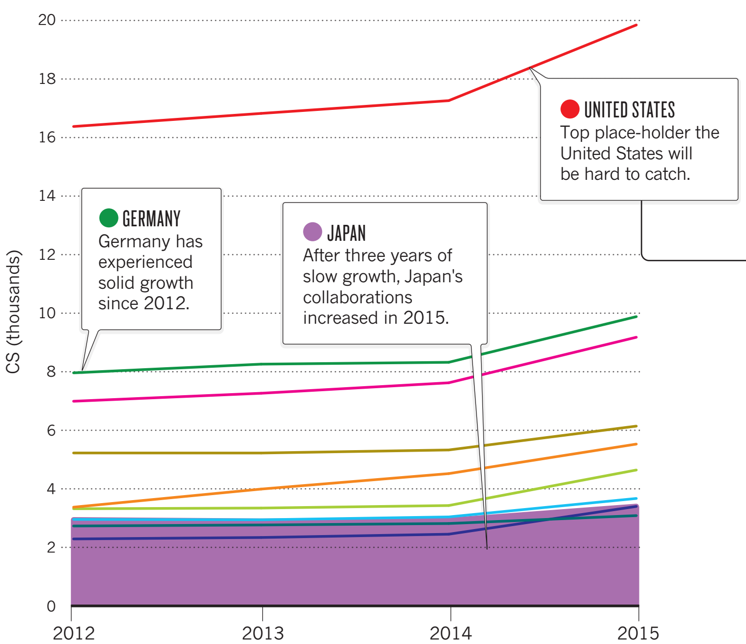
A NETWORK OF KNOWLEDGE

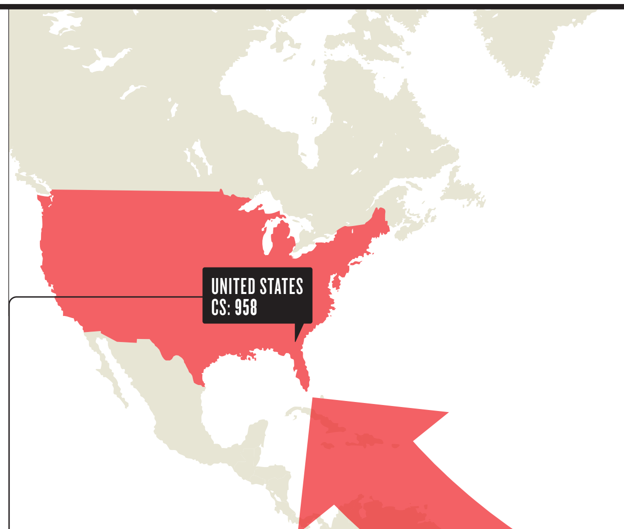
Research breakthroughs are a global pursuit, and while there is healthy rivalry between nations, Japan and its strongest competitors are increasingly joining forces.

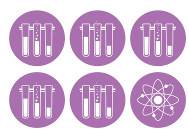
DATA ANALYSIS BY LARISSA KOGLECK

COLLABORATIONS

Japan retained its position as one of the top 10 countries in the Nature Index for its collaborations with domestic and international institutions in 2015. The scope of Japan's partnerships grew between 2014 and 2015, after remaining fairly static since 2012. This phenomenon is not specific to Japan. With the exception of China, which has been growing consistently, most countries showed little movement in their overall output of collaborations between 2012 and 2014. Collaboration score is a measure of the contribution of authors from different institutions to every paper included in the index.







UNITED STATES

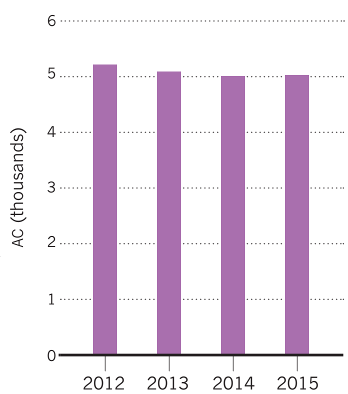
As Japan's top collaborator in the Nature Index, it is little surprise that the USA is also the country with which it has shared the most science Nobel Prizes. The two countries have been joint recipients of six Nobels since 2000, most in chemistry. In 2010 Japanese chemists, Akira Suzuki and Ei-ichi Negishi, the latter having spent most of his career in the USA, shared the chemistry prize with the American, Richard Heck, for their work synthesizing complex carbon molecules widely used in drug manufacturing.

JAPAN BY NUMBERS

ARTICLE COUNT

The number of articles Japanese researchers have published in high-impact journals has remained fairly consistent.

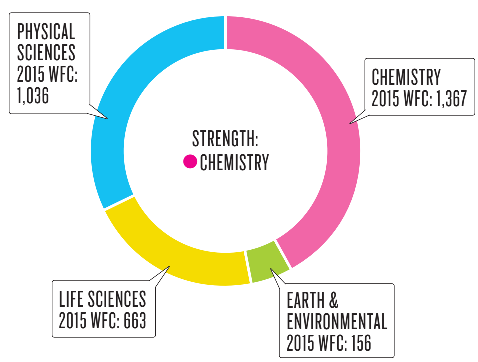
A country or institution's AC is the number of articles in the index that have at least one author from that country or institution.



SUBJECT SPLIT

Chemistry research has been the country's dominant output in the index.

Subjects may overlap. The sum of subject area WFCs may therefore exceed the country's overall WFC.



LEGEND

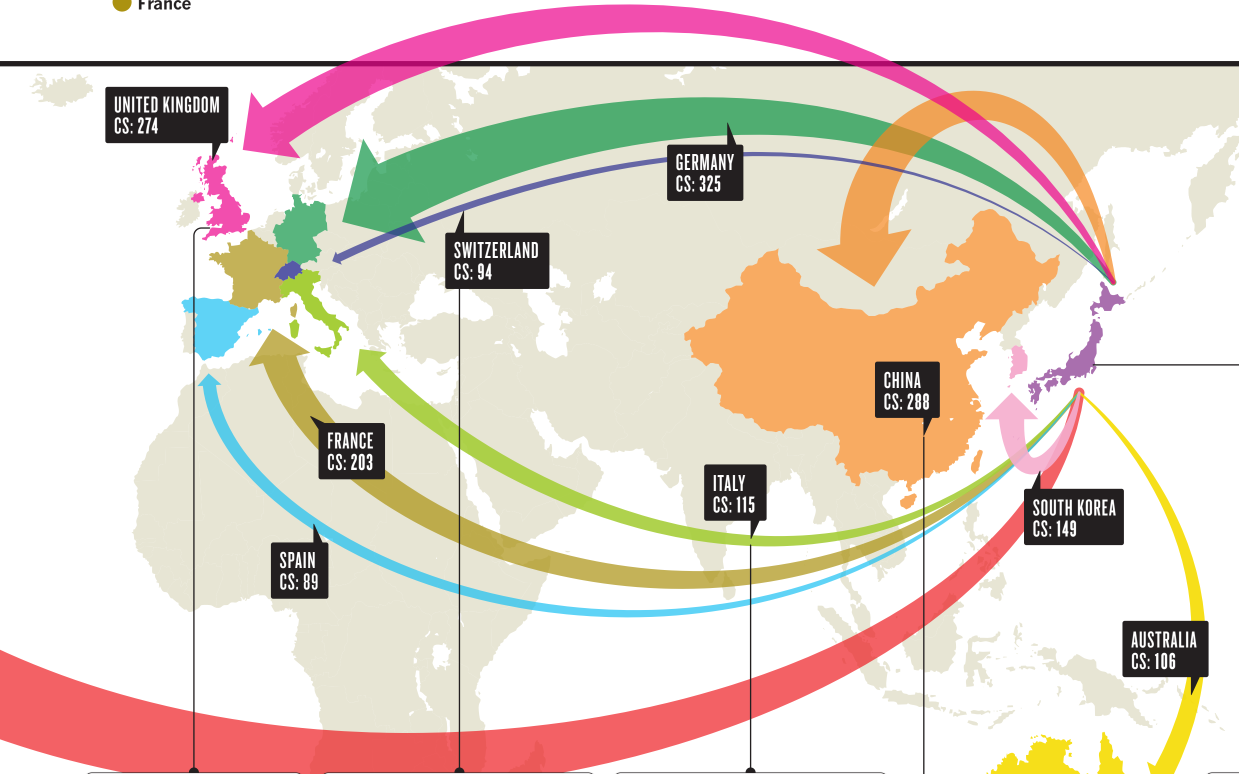
- Japan
- United States
- Germany
- China
- United Kingdom
- France
- South Korea
- Australia
- Italy
- Spain
- Switzerland

Arrow sizes represent the top 10 countries Japan collaborated with in the Nature Index in 2015.

Nobel Prize categories. Each circle represents a prize shared between Japan and another nation since 2000.

Chemistry | Medicine or Physiology | Physics

AC: article count
CS: collaboration score
WFC: weighted fractional count



JAPAN

Japan boasts the second most Nobel Prize Laureates in the natural sciences in the 21st century. Since 2000 15 Japanese scientists have been awarded science Nobels — eight in physics, five in chemistry and two in physiology or medicine — and ten prizes have been shared with scientists from countries that represent some of Japan's strongest collaborations in the Nature Index.

UNITED KINGDOM

The two science Nobels that Japan has shared with the UK since 2000 — in medicine or physiology in 2012 and 2015 — paint a picture of collaboration between the two.

SWITZERLAND

In 2002 Japan received the physics prize and chemist, Koichi Tanaka, shared the chemistry prize with Switzerland's Kurt Wuthrich and the American John Fenn for their development of methods for identifying and analysing biological macromolecules.

ITALY

The single prize shared between Japan and Italy was for discoveries in astrophysics. The Italian Riccardo Giacconi and Japan's Masatoshi Koshiba were awarded the physics prize in 2002, principally for their effort in detecting cosmic neutrinos.

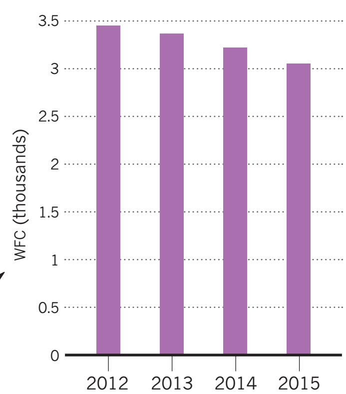
CHINA

In 2015 Japan and China shared their first Nobel Prize. Japanese biochemist, Satoshi Omura, along with his Chinese counterpart, Tu Youyou, were two of three scientists awarded the medicine prize for their discoveries concerning therapies for malaria or infections caused by roundworms.

OUTPUT

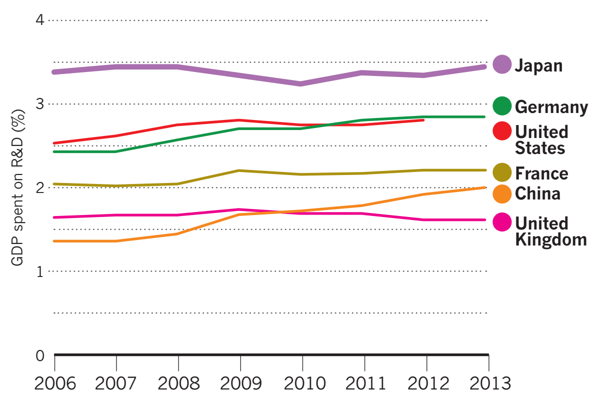
The country's high-quality science output in the Nature Index, measured by WFC, has decreased in the past four years.

WFC apportions credit for each article according to the affiliations of the contributing authors.

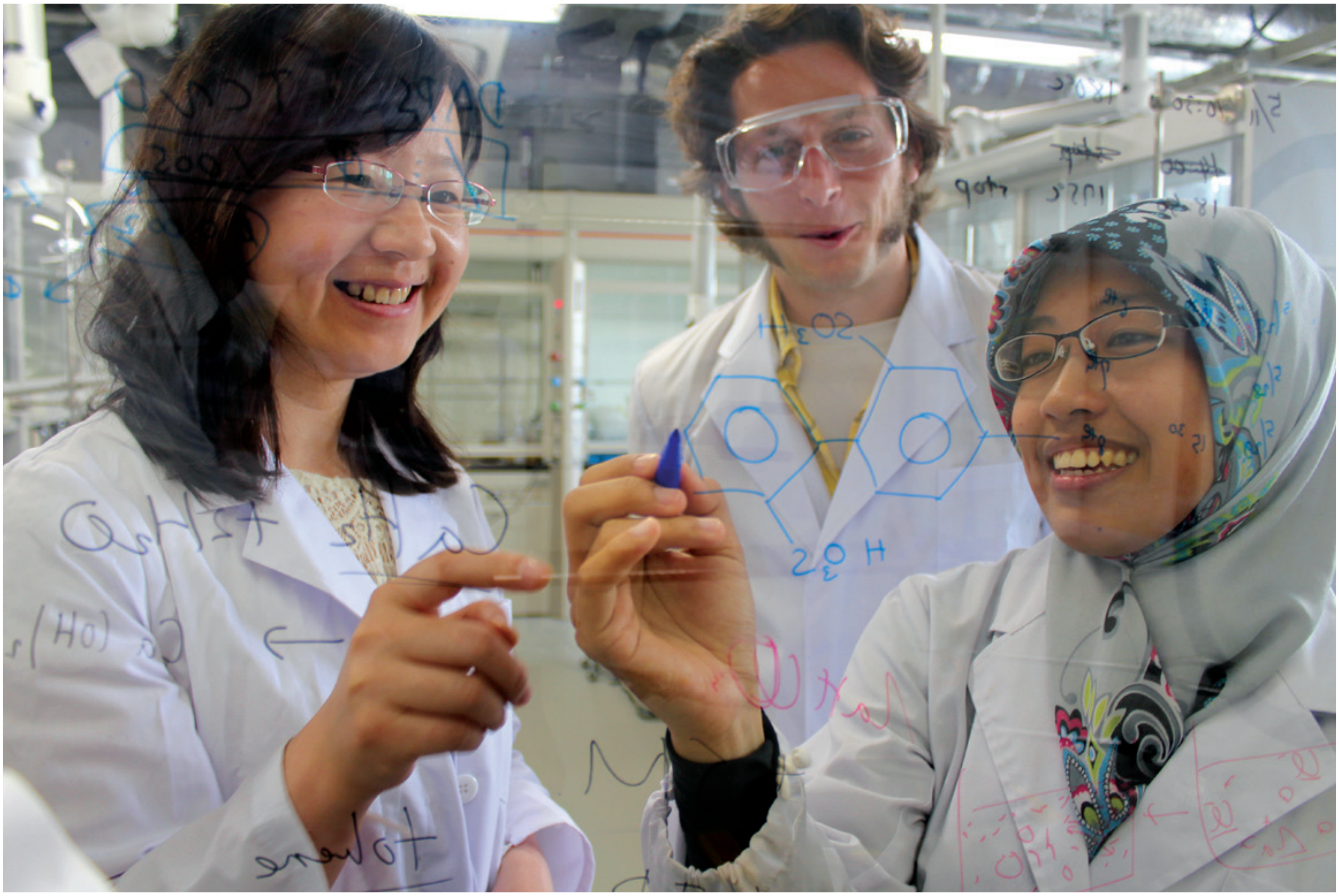


RESEARCH AND DEVELOPMENT

Despite a fall in its science output, Japan has spent a higher proportion of GDP on R&D than its top five collaborators in the Nature Index.



SOURCE: WORLD BANK



Researchers working on electrochemical energy conversion search for fuel cell breakthroughs at the I^{2}CNER at Kyushu University.

FLAGGING FORTUNES INVITE REFORM

Japan's quest to retain its status as a global research leader by enticing foreign students and faculty is the right response, say observers, but efforts are yet to have an impact.

BY ICHIKO FUYUNO

The landlocked former Soviet republic of Uzbekistan seems an odd choice of location for a Japanese education fair, but, in November 2015, representatives from seven leading Japanese universities travelled to Tashkent, the Uzbek capital, for their fifth annual event.

Presented by Nagoya University, it drew hundreds of local students interested in the opportunity to study at prestigious Japanese institutions without having to speak the language.

About 100 Uzbek students come to Japan every year, including about 40 to Nagoya, but Yoshihito Watanabe, the vice president of Nagoya University, would like to attract more.

"Their interest in Japanese education is probably the highest among other countries where we organize similar events," says Watanabe, who is in charge of international education at his university.

From France and the United Kingdom to China, India and Thailand, Nagoya University and other leading Japanese universities hold similar promotional events around the world. They aim to strengthen communication with education officials and local students, and entice them to study in Japan.

This strategy responds partly to a sense among the Japanese research community that its research output and impact has stagnated in recent years. While Japan ranked fifth in the Nature Index in 2015, the country's performance in the database of the world's best natural science research is slipping. Since 2012, the proportion of Japanese contributions to papers in 68 high-impact journals included in the index — its weighted fractional count (WFC) — has fallen 12% (see Top 10 Index performers). The performances of the United States and France have also declined.

"About 100 students from Uzbekistan come to Japan every year, including about 40 to Nagoya."

After decades of keeping universities insular, Japan has realized that diversity and collaboration are essential to producing world-class science and research. To revive Japan's academic competitiveness, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) has been pushing its higher education sector to attract more foreign staff and students. Simultaneously, it has been implementing reforms to make its research environment more competitive with the best research universities in the world. "Internationalization changes the internal structure of universities, which makes them more flexible and on par with global norms," says Petros Sofronis, director of the WPI (World Premier International Research Center Initiative) International Institute for Carbon-Neutral Energy Research (I^{2}CNER) at Kyushu University.

I2CNER

"The Japanese government is doing the right thing," says Sofronis.

But many policy analysts and senior researchers say these efforts aren't enough. They say more reforms are needed to allow research institutions and universities to make decisions about their own structure. "If the government is really determined to raise the level of globalization, current strategies are insufficient," says Hiroshi Nagano, visiting fellow at the Japan Science and Technology Agency, who is a specialist in international science polices.

DECLINING TREND

For more than a decade, the Japanese government has tightened its annual budget for science amid the country's weak economy and growing social welfare demands. The 2016 budget for science and technology is set at 3.45 trillion yen, down 6.4% from its peak in 2012.

The government's R&D investment accounted for 19.5% of the country's total R&D spend in 2013, the lowest among major countries including China, South Korea and the United States.

The president of Suzuka University of Medical Sciences, Nagayasu Toyoda, who observes Japanese science policies closely, says the effect has been "deteriorated support for fundamental research activities, the foundation of our country's strength."

And while the Japanese government, since 2005, has increased the value of competitive grants awarded to scientists for research projects, it has slashed the key funding pool that national universities and research institutions rely on to make themselves more competitive, known as Management Expenses Grants.

It's not just within the Nature Index that Japan's ranking has slipped. According to a recent report by the National Institute of Science and Technology Policy (NISTEP), the number of Japanese papers cited in the top 1% of all journal articles published between 2011 and 2013 fell to seventh in the world, down from fourth a decade ago.

OPENING UP

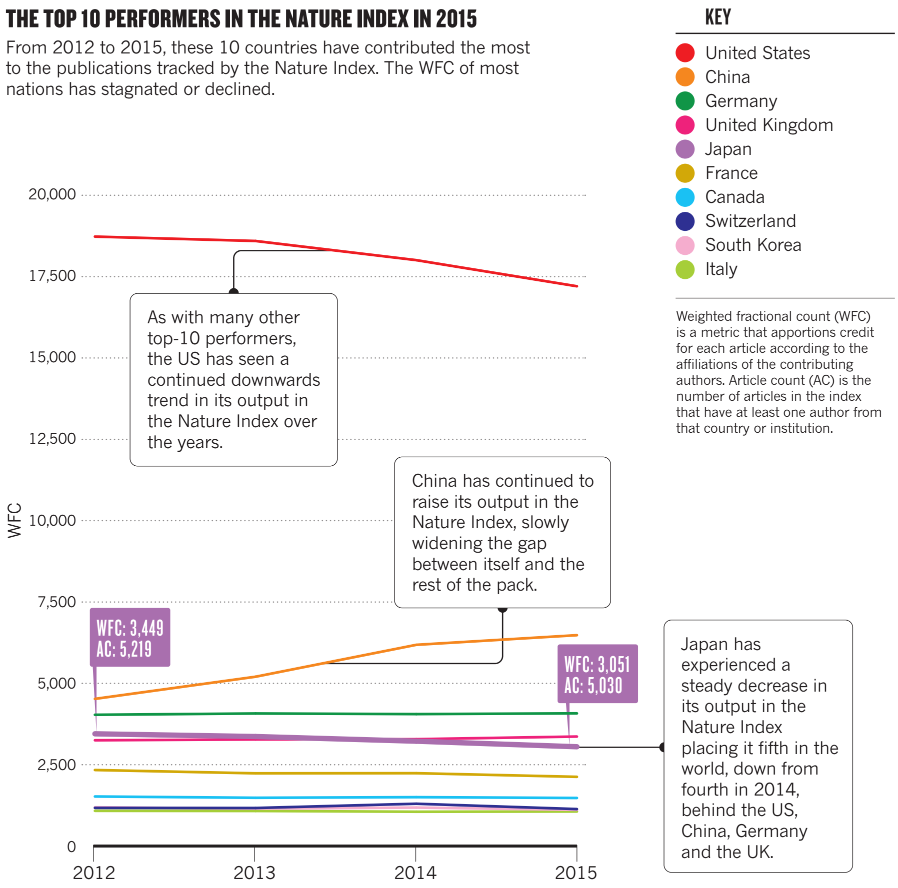
The Japanese government's efforts to open up higher education are not new, but they have accelerated in the past few years. While the budget for science and technology has fallen, the money available for higher education institutions to globalize has surged from 1.9 billion yen in 2006 to 9.34 billion yen in 2016. Between 2009 and 2013, the MEXT established the Global 30 project, funding 13 leading universities to create degree courses in English that would attract foreign students amid intensifying recruitment competition with other countries. As the number of courses taught in English at these universities grew from a handful in 2009 to more than 300 now, the number of overall overseas students grew from 23,083 in 2009 to more than 28,000 by 2013, including

NATSUKI SAKAI/AFLO/ALAMY LIVE NEWS

Prospective students head for an entrance examination at The University of Tokyo's Hongo campus.

THE TOP 10 PERFORMERS IN THE NATURE INDEX IN 2015

From 2012 to 2015, these 10 countries have contributed the most to the publications tracked by the Nature Index. The WFC of most nations has stagnated or declined.



those enrolled at short-term programmes and other non-Global 30 courses. But this modest increase in international students cost Japanese taxpayers 14.7 billion yen over five years, and by 2013 the country's percentage of foreign students, 3.4%, remained well below the OECD average of 9%. The increase in foreign students is also yet to be reflected in the output of high-quality research for most Global 30 participants (see below).

NEXT STEP

While the Global 30 project succeeded in increasing the number of university courses offered in English, the MEXT launched another programme in 2014 to catapult more Japanese universities into the world's top 100 institutions. The Top Global University Project, often referred to as the Super Global University programme, funds 37 universities to change their organizational structures, partly to deepen ties with international institutes. Thirteen 'Type A' universities — many of which are among the top ten institutions in the Nature Index in Japan in 2015 — will receive up to 500 million yen a year until 2023 to become a top 100 university in rankings such as the Times Higher Education list. Many Type A members are striving to create joint degrees with top-level overseas institutions. In April 2015, Tohoku University established the world's first graduate programme in spintronics in collaboration with Johannes Gutenberg-Universität Mainz. The spintronics course fits with Japan's strength in physical sciences, evidence of which is seen in the Nature Index (see Subject Split, S102).

Another 24 'Type B' institutions will be given up to 300 million yen to use their unique strengths to contribute to local economies.

Perhaps Japan's most successful programme to promote internationalization is the WPI, which launched in 2007 and comprises nine institutes. The official language of these centres is English and an average 40% of researchers are from abroad. WPI institutes accounted for 4.63% of the world's most highly cited papers between 2007 and 2013, the third largest contribution after the Rockefeller University and the Massachusetts Institute of Technology (MIT).

> ***"We have to create new measures to assess how universities bring competitive personnel into the global society."***

But the executive director of RIKEN, Yoichiro Matsumoto, says there are few indicators to assess the success of these programmes. "The government evaluates the impact of programmes like Global 30 only by numbers. But we have to create new measures to assess how universities bring competitive personnel into the global society," says Matsumoto.

FUNDAMENTAL PROBLEMS

Charles Yokoyama, director of research administration at the RIKEN's Brain Science Institute (BSI), says that despite Japan's efforts to attract overseas staff and students, it's hard for the country to keep up.

"Globalization activities are slower than they can or should be to keep up with global competition," he says. The RIKEN BSI is one of the most international institutions in Japan yet only a few postdocs are trained, or awarded faculty positions at elite US universities. The pool of talented researchers largely bypasses Japan, says Yokoyama.

Hiring excellent researchers is key to attracting top-level students from abroad, but "Japan can be no match for the United States and other major countries so far in terms of the level of salaries and welfare packages for families," adds Takaaki Kajita, director of the Institute for Cosmic Ray Research at the University of Tokyo.

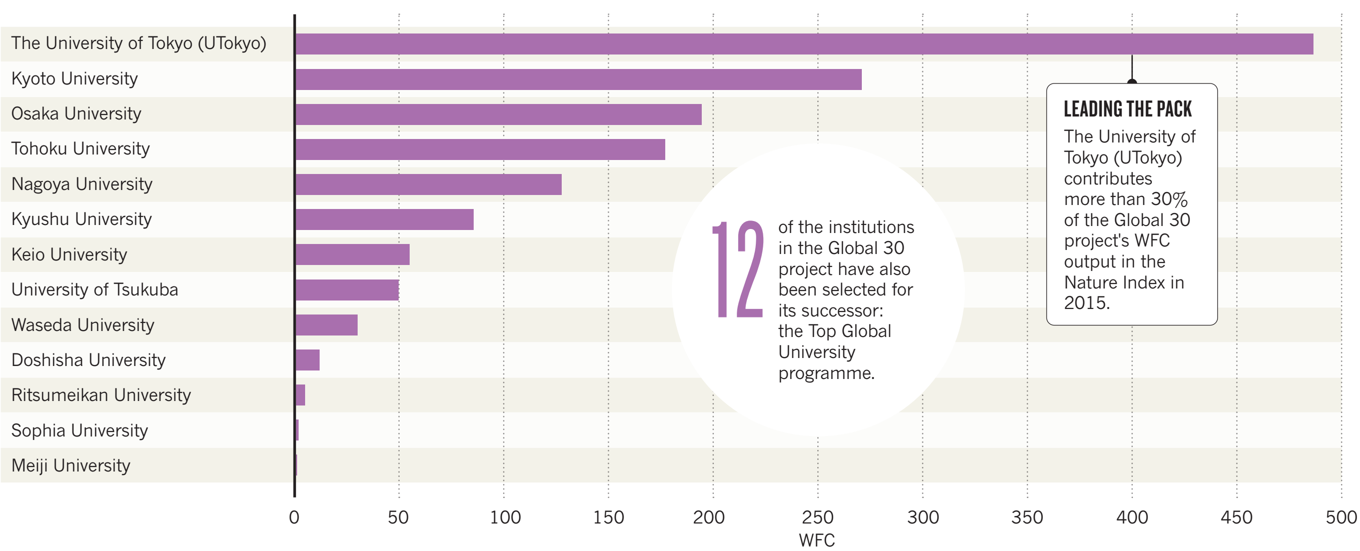
While Japan may struggle to compete with other nations in terms of recruitment, Japanese researchers are reaching out to partner with their global peers. Between 2014 and 2015, Japan's collaborations with China, the United Kingdom, France and Germany grew. Its partnerships with the United States also increased, an upsurge to a three-year decline (see Japan's top partnerships).

Shizuo Akira, director of the WPI Immunology Frontier Research Center at Osaka University, says many young Japanese researchers are unwilling to move labs or go abroad for postdoc positions and miss opportunities to learn from different leaders. Akira, one of the world's most cited immunologists, says young Japanese researchers are reluctant to move as competition for positions in Japan is intense after returning from abroad. "Unless we establish a practical career path that presents young fellows as visibly prominent scientists on the global stage, there is little hope for the future of Japanese science."

The government is aware of these issues and is implementing various measures to strengthen ties with overseas collaborators.

GLOBAL 30 PROJECT UNIVERSITIES IN THE NATURE INDEX

The MEXT Global 30 project funded 13 leading universities to create degree courses in English that would attract foreign students. The courses surged from a handful in 2009 to more than 300 now.



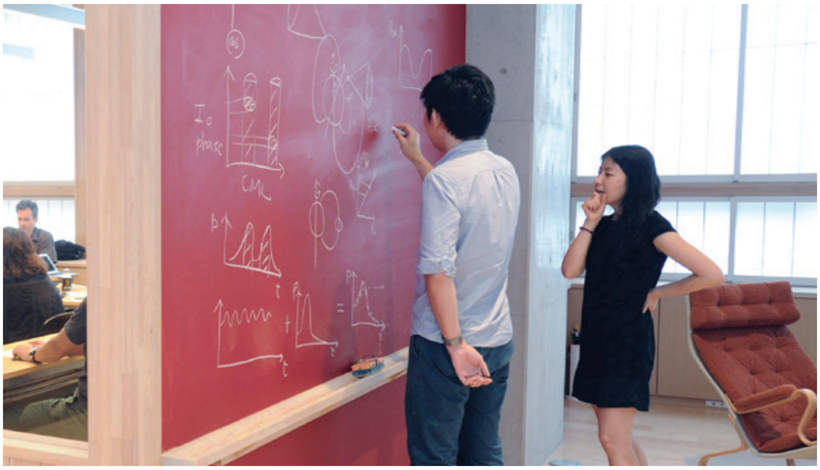
The MEXT is in the process of revising Japan's biggest competitive grant, the Grants-in-Aid for Scientific Research, so it can provide greater support for joint international research. In addition, the ministry set aside 5.6 billion yen this year to accelerate intellectual exchange, allowing about 520 young Japanese researchers to work abroad for two years, and 1,100 overseas researchers to come to Japan.

But Nagano worries these measures are relatively small due to the government's tight science budget, and many have not led to a rise in Japan's rankings. "If Japan wants to raise its global ranking, we need to recruit good students. To do that, we need good faculty. But there are no measures that combine these two points," he says.

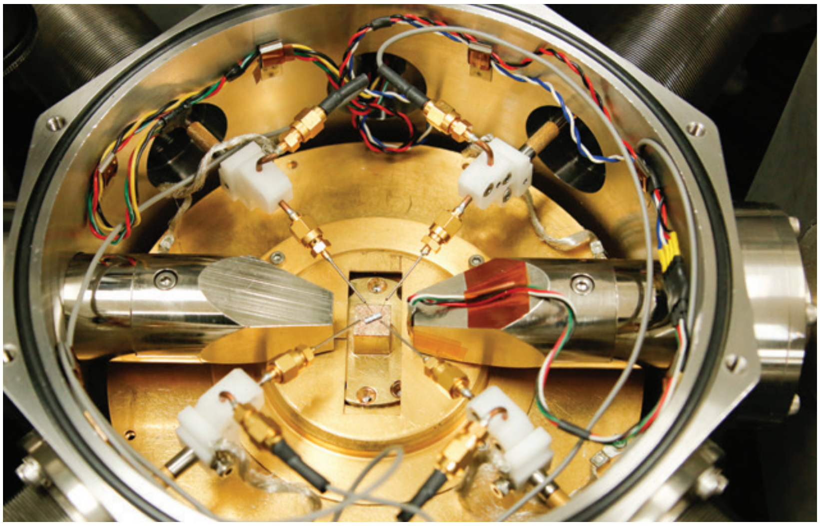
Meanwhile, Yokoyama observes that Japan's traditional solutions to globalize research often focus on the recruitment of international visitors such as students and junior faculty, but more basic reforms of domestic graduate programmes are needed for Japan to develop a globally competitive STEM (science, technology, engineering and mathematics) workforce.

"The level, approach and pace of globalization of Japan's graduate programmes have fallen short of elite global standards including the key issue of training in high-level English communication," he says.

Atsushi Sunami, the government's advisor for science policies at the National Graduate Institute for Policy Studies, says comprehensive reforms of national universities are currently underway to reorganize faculties, strengthen governance and review funding and personnel systems, but their effect will take time. ■



TOKYO TECH

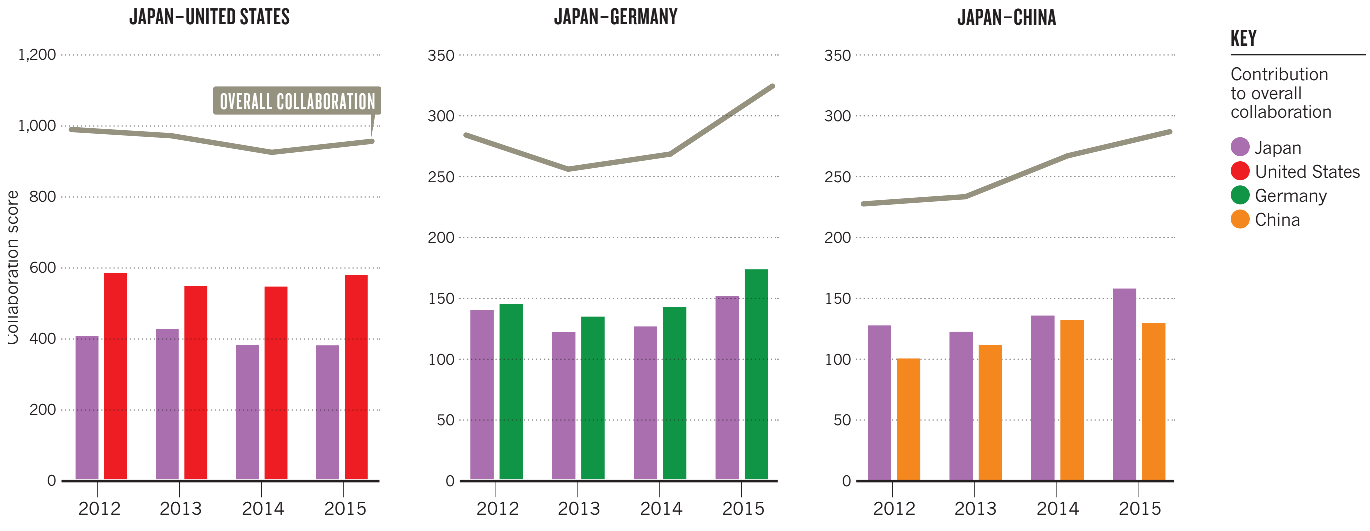


TOHOKU UNIVERSITY

Top: Students work on a solution at the Earth Life Science Institute, a designated WPI centre. Bottom: A system for testing the properties of spintronics components at Tohoku University.

JAPAN'S TOP PARTNERSHIPS

Japan's most prolific collaboration partnerships are with the United States, Germany and China. The largest collaboration with the United States has seen a slight decline, while partnerships with Germany and China have grown.



The Kamioka Gravitational Wave Detector (KAGRA), housed in a giant tunnel, is a project of University of Tokyo's Institute for Cosmic Ray Research.

EYE ON THE PRIZE, BUT GRASP LOOSENS

In most of the past 15 years, Japanese scientists have been named Nobel Laureates for work in chemistry and physics, but illustrious awards may be harder to come by in the next era.

BY TIM HORNYAK

Modern Japanese research is a tale of two eras. While the nation's scientists have been recognized with Nobel Prizes for work in fields such as chemistry and physics in the majority of the last 15 years, some observers predict that such awards will be more elusive for the next 15.

University of Tokyo physicist, Takaaki Kajita, the most recent laureate, was honoured, along with Canada's Arthur B. McDonald, in 2015 for showing that neutrinos can oscillate, demonstrating that they also have mass — a revelation that defied the standard model of particle physics.

The achievement solidified the University of Tokyo's dominant position over other Japanese research centres. It holds the country's top spot in 2015 in the Nature Index, a database of the contributions scientists have made to 68 high-quality science journals. The index tallies the annual number of scholarly papers affiliated with a particular institution, the relative contribution of each author to a particular study, known as the fractional count, and the weighted fractional count (WFC), which adjusts for the relative abundance of papers in astronomy and astrophysics. Between 2014 and 2015, the University of Tokyo's WFC rose 6%, the largest increase of any institution in the Nature Index top 10.

Perhaps unsurprisingly, seven of the top 10 Japanese institutions in the index form the state-backed National Seven Universities, a group analogous with the Ivy League schools in the United States.

The top 10 carry out research supported by a combination of government funding and collaborations with the private sector, though the latter has been taking on an increasingly significant role.

Despite the country's impressive Nobel Prize cabinet, by other measures of success there may be cause for concern. Between 2012 and 2015 the output of high-quality science counted in the Nature Index declined for eight out of Japan's top 10 institutions. This follows a wider trend within Japanese institutions where the overall number of papers published in five natural science disciplines has fallen in recent years, according to an analysis by Nagayasu Toyoda, president of Suzuka University of Medical Sciences.

> ***"The number of scientific articles in Japan normalized by population or GDP is less than half that of many OECD countries."***

"The number of scientific articles in Japan normalized by population or GDP is less than half that of many OECD countries," says Toyoda, who anticipates fewer Nobel Prizes for Japan.

"The main reason for the decline in research activity at Japanese universities is budget cuts to basic subsidies to national universities by about 1% or more per year during the past 10 years and beyond."

This will have an impact on Japanese science, as top researchers have historically been able to leverage their reputation to try to break new ground. For instance, Kajita has been working toward the direct detection of a gravitational wave. Predicted by Einstein's theory of relativity, gravitational waves can distort space-time in minute quantities — equivalent to the diameter of a hydrogen atom for the distance between the Earth and the Sun.

The recent detection of gravitational waves could be the beginning of a new kind of astronomy. Kajita is supervising a project to build a 3-km laser interferometer. "The beginning of the universe is an ultimate goal, but for the moment our hope is to observe the formation of black holes."

THE ASAHI SHIMBUN/GETTY IMAGES

TOP OF THE NATION

Japan's top 10 overall performers in the Nature Index in 2015 include the National Seven Universities, government funded research institutions and up-and-coming research universities.

1 **UNIVERSITY OF TOKYO**
2015 WFC: 486

2 **KYOTO UNIVERSITY**
2015 WFC: 270

3 **OSAKA UNIVERSITY**
2015 WFC: 194

4 **TOHOKU UNIVERSITY**
2015 WFC: 176

5 **RIKEN**
2015 WFC: 137

See RIKEN profile on S128

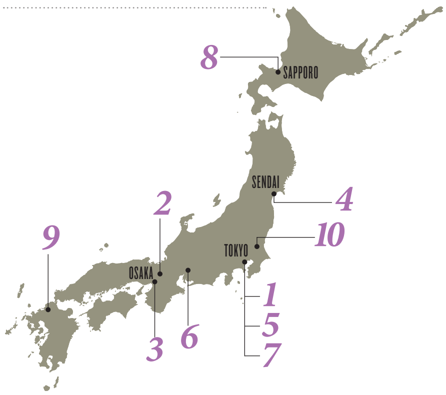
6 **NAGOYA UNIVERSITY**
2015 WFC: 127

7 **TOKYO INSTITUTE OF TECHNOLOGY**
2015 WFC: 120

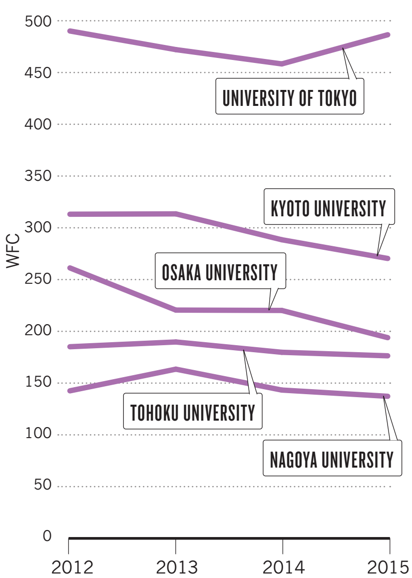
8 **HOKKAIDO UNIVERSITY**
2015 WFC: 103

9 **KYUSHU UNIVERSITY**
2015 WFC: 85

10 **NATIONAL INSTITUTE FOR MATERIAL SCIENCE**
2015 WFC: 82



WFC output from 2012 to 2015 for Japan's top 5 universities in the Nature Index.



JTB PHOTO/UIG/GETTY IMAGES

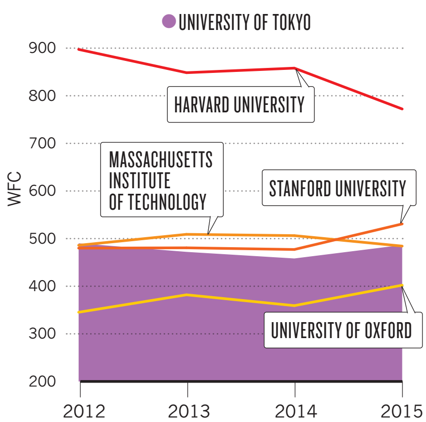
The University of Tokyo has been a stronghold of excellence in Japanese science since its establishment.

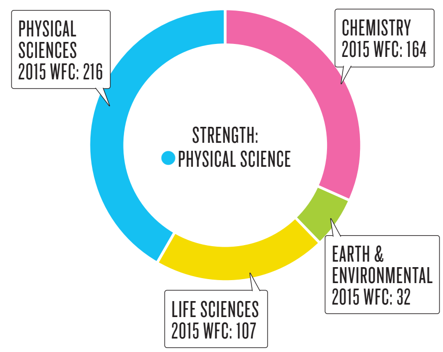
UNIVERSITY OF TOKYO

2015 WFC: 486 **2015 AC:** 1,377

2015 NATURE INDEX ACADEMIC RANKING: 3

Measuring the performance of University of Tokyo in a global context over time. The line graph below compares University of Tokyo (purple) with peers with a similar WFC output in the Nature Index 2015. The donut shows WFC output by subject area in 2015.





•Subjects may overlap. The sum of subject area WFCs may therefore exceed the institution's overall WFC.

Founded in the decade after Japan's emergence from feudalism in 1868, the University of Tokyo has long presided over the ranks of Japanese centres of learning. Today, it comprises 10 faculties, 15 graduate schools, 13 research centers and facilities located throughout Japan. Among its scholars are eight Nobel Prize winners and one Fields Medal winner. To maintain its research output, the university relies on research grants and collaborative research with industry for about one-third of its budget. Of its 23.5 billion Grants-in-Aid for Scientific Research from the state-backed Japan Society for the Promotion of Science (JSPS), more than half is used to support fundamental research, and 16% goes to support young researchers.

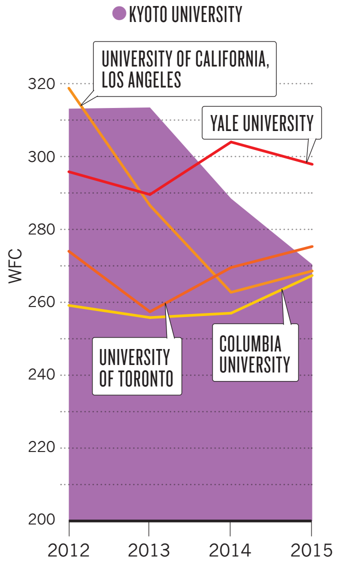
"Members of the University of Tokyo are active across the full spectrum of human endeavour."

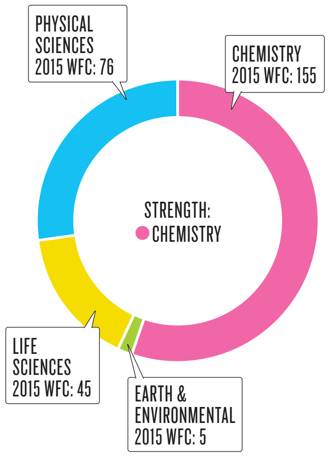
"Research today is ever more specialized and increasingly requires the efforts of teams working across disciplinary boundaries," says Kazuo Hotate, executive vice president for research at the university. "One of our great strengths is that members of the University of Tokyo are active across the full spectrum of human endeavour, providing immense opportunity for interdisciplinary collaboration both within the university and with our partners across Japan and the world."

KYOTO UNIVERSITY

2015 WFC: 270 2015 AC: 715

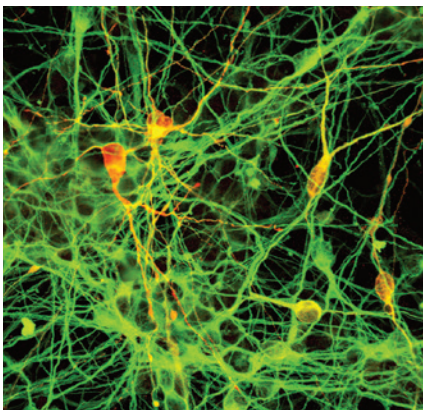
2015 NATURE INDEX ACADEMIC RANKING: 14





•Subjects may overlap. The sum of subject area WFCs may therefore exceed the institution's overall WFC.

Like the University of Tokyo, Kyoto University has produced a number of outstanding researchers since its establishment in the late 19th century — nine have won a Nobel Prize, including Shinya Yamanaka in 2012 for discovering that mature cells can be reprogrammed into so-called induced pluripotent (iPS) cells (described in the journal *Cell*). Yamanaka wrote about streamlining iPS cell research in *Nature* in 2009. Today, Kyoto has 14 research institutes including the Center for iPS Research and Application (CiRA), where Yamanaka is director. Following partial privatization in 2004, along with all other national universities, government block funding has been decreasing. Kyoto University has responded by increasing joint projects with corporate partners as well as taking on corporate-sponsored research, which now accounts for about 42% of the research budget. An example of the embrace of industry is T-CiRA, a 10-year collaboration between CiRA and Takeda Pharmaceutical Company, Japan's largest drug maker, to develop clinical applications of iPS cells. "This system of innovation is seen as being in tune with the city of Kyoto, where tradition combines with creativity to result in an environment fostering venture businesses," says the university spokesman, David Hajime Kornhauser. "KU has tended to attract the best minds in Japanese academia: those who are seeking a place to pursue their ideas freely and express their creative spirit to the utmost".



Neurons derived from iPS cells, whose discovery earned Shinya Yamanaka a Nobel Prize in 2012.

ASUKA MORIZANE/CIRA/KYOTO UNIVERSITY

"KU has tended to attract the best minds in Japanese academia."

OSAKA UNIVERSITY

2015 WFC: 193 2015 AC: 532

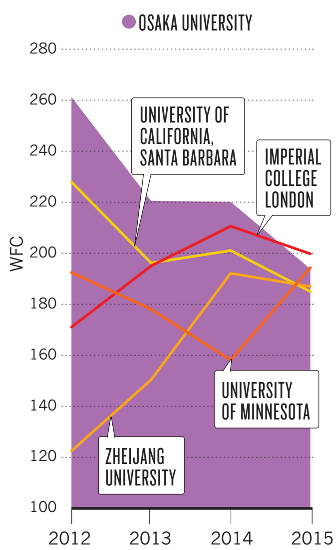
2015 NATURE INDEX ACADEMIC RANKING: 34

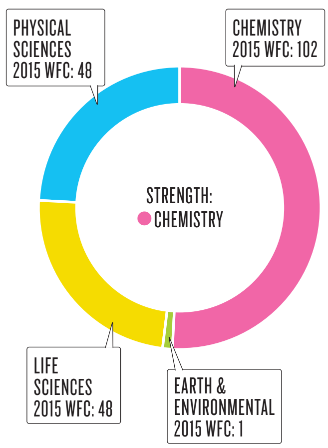
JUN SAKANOUE

Shizuo Akira, a leader in the field of immunology.

Osaka University is known for its immunology discoveries. It was at this institution that the emeritus professor, Tadamitsu Kishimoto, identified the immune protein interleukin-6 in the 1980s. Today the university is trying to clear a path to the next breakthroughs in the field through the Osaka University Immunology Frontier Research Center (IFReC). Opened in 2007 as one of the Japanese government's "globally visible" World Premier Institute Research Institutes, IFReC aims to broaden the scope of immunology research to include study of cells in the body and come up with solutions to immune diseases including vaccine development. The staff of more than 20 researchers in immunology, bioinformatics and imaging have elucidated links between the immune system and metabolic syndromes including lipolysis and gout, described in 2013. Director, Shizuo Akira, said the centre had published more than 1,000 papers on a wide variety of immunology fields, almost 10% of them in high-impact journals. Akira, who studied in Kishimoto's lab, said: "Our laboratory is actively involved in the determination of new subsets of macrophage and regulation of post RNA transcription. In 2016, I expect robust achievements in these fields."

"In 2016, I expect robust achievements in these fields."





•Subjects may overlap. The sum of subject area WFCs may therefore exceed the institution's overall WFC.

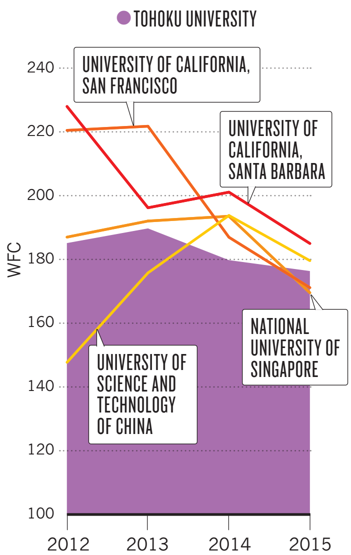
TOHOKU UNIVERSITY

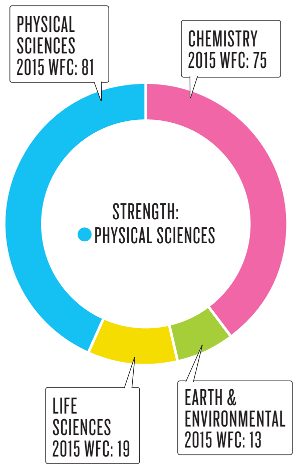
2015 WFC: 176 2015 AC: 429

2015 NATURE INDEX ACADEMIC RANKING: 38

TOHOKU UNIVERSITY

Hideo Ohno uses a stereomicroscope in the study of spintronics at Tohoku University.





•Subjects may overlap. The sum of subject area WFCs may therefore exceed the institution's overall WFC.

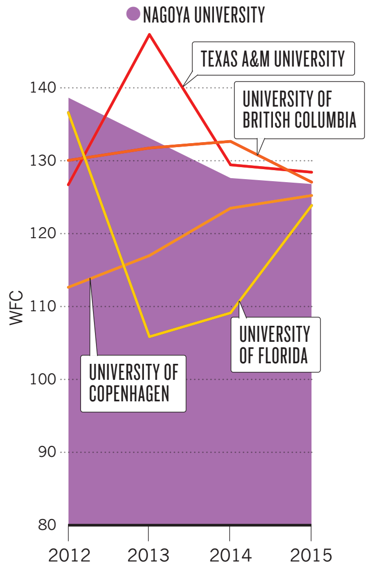
Hideo Ohno spent 10 years researching semiconductors in his hometown of Sapporo before he became curious about the possibilities of combining magnets with these indispensable devices of the information age. In 1994 he began to explore this in earnest, when he joined the faculty of Tohoku University's Department of Electronic Engineering. Ohno is now the director of the Center for Spintronics Integrated Systems. "Spintronics is a broad science and engineering field, in which we utilize the 'spin' (a magnetic property) of the electron," says Ohno. In a 2015 *Nature Physics* paper, Ohno and collaborators described the so-called 'creep' slow motion of domain walls in ferromagnets. As a practical application, spintronics could transform household appliances and factory equipment into 'smart' devices that collect and exchange data via the internet. It has been used to make high-performance nonvolatile memory storage, which doesn't require energy to retain information. This helps countless processors and controllers stand by without power, an essential quality for the 'Internet of things'. Joining Tohoku gave Ohno the opportunity to lead a group, and access the best equipment to develop new materials and study spintronics. So what does Ohno have in store for 2016? "This is top secret," he says.

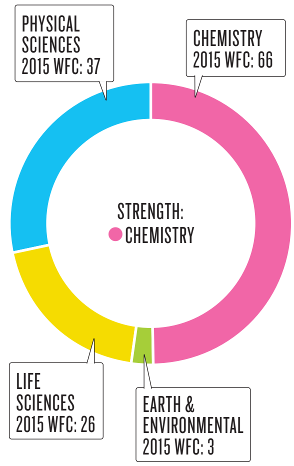
AP/PRESS ASSOCIATION IMAGES

NAGOYA UNIVERSITY

2015 WFC: 126 2015 AC: 433

2015 NATURE INDEX ACADEMIC RANKING: 60





•Subjects may overlap. The sum of subject area WFCs may therefore exceed the institution's overall WFC.

Like every kid growing up in Hamamatsu City, Hiroshi Amano knew about hometown legend Kenjiro Takayanagi, a high-school teacher and inventor who created the first fully electronic TV screen in 1926. But Amano also dreamed of exceeding his achievement. The spark came when he studied engineering at Nagoya University and became interested in PCs. He wondered whether compact light-emitting diodes (LEDs) could replace the bulky cathode-ray tubes that were used in Takayanagi's early TVs as well as PC displays. "I believed that if I could develop a blue LED, I would change the world," Amano says. "Of course, I was a fledgling student and I knew nothing about the difficulty of the subject." His boldness paid off. In 1985, Amano produced the world's first high-luminance blue LED, along with his supervisor, Isamu Akasaki, who focused on gallium nitride as a promising material. His research has also appeared in *Nature Materials*. The achievement won them the 2014 Nobel Prize in Physics, shared with Shuji Nakamura. Today blue LEDs are used in everything from mobile phone backlights to energy-efficient traffic signals and lights for plant cultivation.

Blue LEDs illuminate the Nagoya TV tower to celebrate their launch.

Nagoya is a national university that traces its roots back to a medical school founded in 1871 in a samurai court. These days the university's strengths are chemistry and physical sciences.

With a relatively light teaching load compared to faculty at private universities, Amano says he can concentrate on his research which is now focused on the po10tial of deep ultraviolet (DUV) LEDs for sterilization. "We could provide safe water for 2.4 billion people without safe sanitation and 663 million people who cannot access safe drinking water," he says. ■

SWIMMING AGAINST THE TIDE

The rising stars on Japan's research landscape include traditional big-hitters and newer institutions selected for initiatives to boost their global standing and promote autonomy.

BY MARK ZASTROW

Institutions both public and private are counted among Japan's rising stars — those that have gained the most in the Nature Index since 2012 in terms of the absolute or relative increase in their weighted fractional count (WFC).

Four of the index's top 10 institutions are part of the country's flagship national university system, including three of the top four. The rest are split between governmental or semi-governmental research institutions. and fully private universities.

Many of the institutions have been selected for various government programmes intended to boost their global standing and internationalize their faculty. The top two, the Tokyo Institute of Technology (Tokyo Tech) and Okayama University, are members of the Top Global University Project, often referred to as the Super Global Universities programme, which started in 2014. And Doshisha University, a private institution in Kyoto, was also one of the beneficiaries of that programme's forerunner, the Global 30.

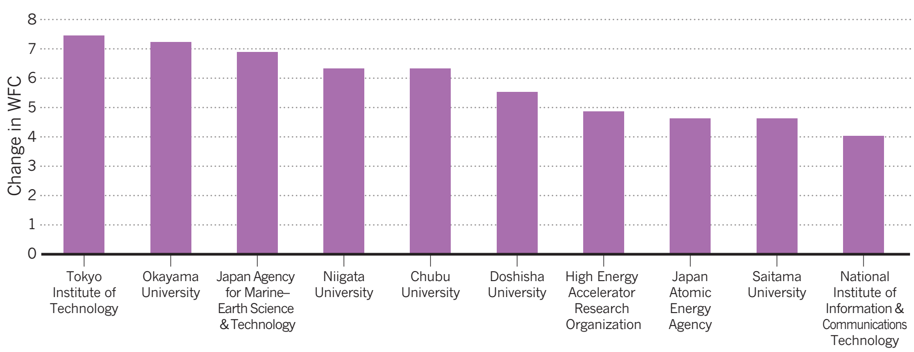
Tokyo Tech also hosts the Earth-Life Science Institute, which focuses on fundamental research on the origins of life on Earth and is one of nine high-profile centres funded by the government's World Premier International Research Center (WPI) Initiative.

These initiatives can be seen as part of the Japanese government policy that holds that "universities and institutes should have their own individuality," says Kumi Okuwada, a senior fellow at the National Institute of Science and Technology Policy (NISTEP). However, the rise of these institutions takes place against a backdrop of concern that the nation's research funding and output is stagnating.

Nevertheless, government-backed institutions

TOP 10 RISERS

These 10 institutions have shown the largest increase in WFC in the Nature Index from 2012 to 2015.



JAMSTEC
JAMSTEC'S *Chikyu*, a giant of deep-sea drilling, was at the forefront of investigations into the fault that caused the 2011 Tohoku earthquake.

are well-represented among those which have an increasing high-quality research output. One such organization is the Japan Agency for Marine-Earth Science and Technology (JAMSTEC). With headquarters in Yokosuka, south of Yokohama, it maintains offices on both sides of the Pacific and operates more than a dozen research ships and remotely operated vehicles. That includes the drilling vessel *Chikyu*. In 2012 it set a then-world record for the deepest hole in deep-sea drilling and, in the aftermath of the 2011 Tohoku earthquake and tsunami, investigated the fault that caused it.

Another government organization rising in the rankings is the Japan Atomic Energy Agency, the nation's nuclear R&D organization. Based in Tokai, north of Tokyo, it operates more than a dozen research institutes, employing a staff of more than 4,000. In addition to its ongoing nuclear power research (including participation in the international ITER fusion reactor), it has devoted resources in recent years to the containment and clean-up of the Fukushima Daiichi reactor following the 2011 quake.

The other government research institute in the top ten — the High Energy Accelerator Research Organization (KEK)—can count a Nobel Laureate among its recent faculty: Makoto Kobayashi received the physics prize in 2008 for his work on symmetry violation in particle physics. Only one other private university made the list—Chubu University near Nagoya, which has colleges in biotechnology, health sciences, and engineering.

TOKYO INSTITUTE OF TECHNOLOGY

2015 WFC: 120 **2015 AC:** 365

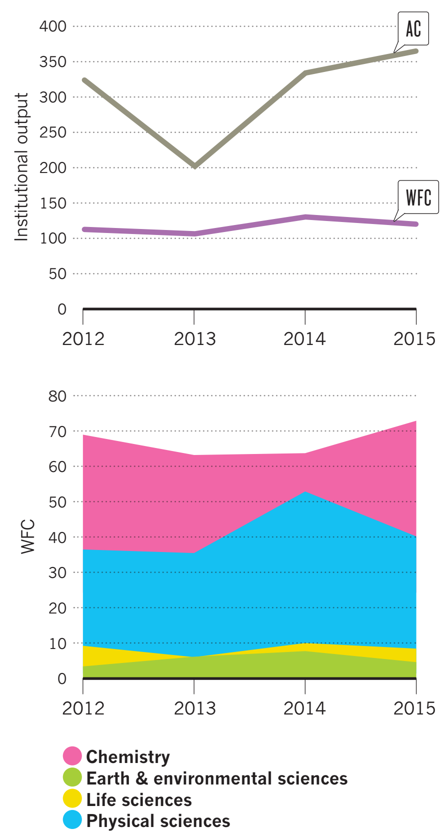
Japan's fastest-growing research institute in the Nature Index is also the largest technical university: the Tokyo Institute of Technology. Its WFC increased from 112 in 2012 to 120 in 2015. More than 40% of its revenue comes from public funds, and it has a research budget for commissioned projects in 2015–16 of 13.8 billion yen (US$123 million).

Tokyo Tech's executive vice president of research, Makoto Ando, credits the university's comprehensive range of research interests as key to its rise in the Nature Index, citing research in materials science, cell biology, and fibre optic communication — as well as the Tsubame supercomputer, which the university owns and operates. Tsubame is the world's second-most energy efficient supercomputer, one of the fastest owned by a university, and has powered compelling research ranging from measuring blood flow in drug simulations,to modelling seismic activity of the volatile Nankai Trough off the island of Honshu.

The university operates 16 research institutes, laboratories across its three campuses situated in the capital city and surrounds. Ando says reforms from April 2016 will include a focus on globalization: the university has nearly 200 visiting international researchers, 10% of its students come from outside Japan, and it hopes to further boost those numbers.

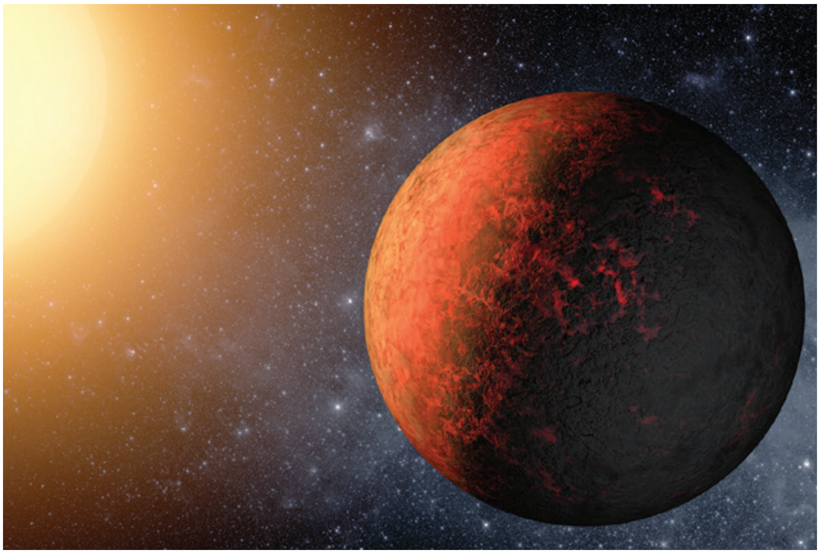
PHYSICAL STRENGTH

The line graph shows overall trends for both WFC and article count (AC) since 2012. The area plot highlights which subject has contributed most to the increase in WFC.



TOKYO TECH

NASA/AMES/JPL-CALTECH



JTB PHOTO/UIG/ GETTY IMAGES

Top: The Tsubame supercomputer at Tokyo Tech has powered research ranging from blood flow in drug simulations to modelling seismic activity.
Middle: An impression of the exoplanet Kepler-20e. These molten-covered planets are the subject of work by geophysicists at Okayama University.
Bottom: Sado, off the coast of Honshu is the site of Niigata University's marine biology station.

OKAYAMA UNIVERSITY

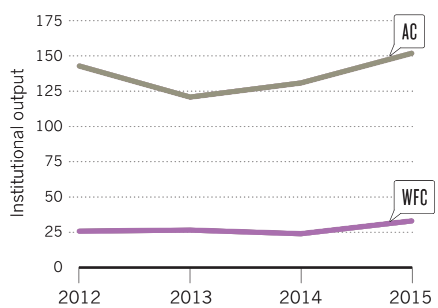
2015 WFC: 33 2015 AC: 152

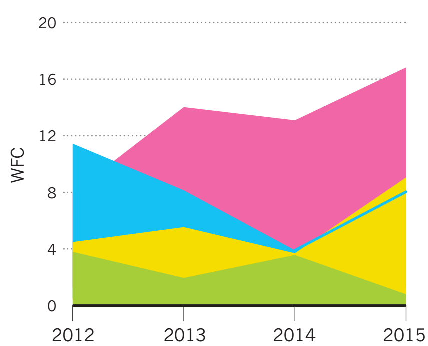
Okayama University's rise in the Nature Index is the second-largest in Japan, making it also the overall second-most prolific institute on the list. Its vice president of research, Shinichi Yamamoto, credits this to a focused effort to selectively fund distinctive research projects, helped by funding from the nation's science ministry, which allowed it to create an organization for interdisciplinary studies.

One of Okayama's strengths is in physics, from high-energy physics to its geophysical research institute. The latter's expertise is also being applied to the study of other worlds. For example, a recent study in *The Astrophysical Journal* coauthored by Okayama's George Hashimoto proposed a new target for astronomers hunting for exoplanets: planets with an entirely molten surface, covered with magma oceans. Scientists expect that most rocky planets, including Earth-like ones, go through this phase early in their evolution. The team's models showed that they cool significantly after a million years or less, rendering their heat signatures practically invisible. However, the steam rising from the still-warm magma would be highly reflective, making them promising targets for the upcoming generation of telescopes.

Okayama's other major research centre focuses on plant science and photosynthesis. Yamamoto says the university has plans to increase its efforts in medical research, cybersecurity and big data.

CHEMISTRY CHAMPS





JAPAN AGENCY FOR MARINE-EARTH SCIENCE AND TECHNOLOGY

2015 WFC: 22 **2015 AC:** 84

Below the abyss, there is the hadal, as in Hades of the Greek underworld. This is the realm of the deepest sea trenches. "It is the least explored biosphere on Earth," says environmental microbiologist, Takuro Nunoura, of the Japan Agency for Marine-Earth Science and Technology (JAMSTEC).

In the ocean's deep abyssal zone, food is scarce, and microbes eke out what energy they can mostly from rocks. The hadal is even less surveyed, and previous work had been conducted like classical species hunting. But in a 2015 study published in the *US Proceedings of the National Academy of Sciences*, Nunoura and colleagues brought to bear the techniques of modern environmental microbiology.

To his surprise, they found a thriving ecosystem of microbes living off organic matter at the bottom of the Mariana Trench. He suspects that organic matter collected on the trench's slopes gets churned into the water during underwater earthquakes, nourishing this unique ecosystem.

JAMSTEC, which has been the nation's premiere oceanographic institute since its founding in 1971, rose in the index by 47% between 2012 and 2015. Nunoura says this coincides with the retirement of the institute's first generation of scientists, and the rise of younger scientists to leadership positions.

NIIGATA UNIVERSITY

2015 WFC: 13 **2015 AC:** 53

The port of Niigata is the largest city on the Sea of Japan — the more sparsely populated of the nation's coasts—and Niigata University is one of the coast's largest national universities with more than 10,000 undergraduates and 2,200 graduate students.

Not surprisingly, much of its research reflects regional interests or challenges. It operates a marine biological station on Sado Island, 50-km offshore.

One of its major research centres, the Research Institute for Natural Hazards and Disaster Recovery, plays a key role in assessing earthquake and tsunami risk and preparations along the Sea of Japan coast and in the mountains just inland. It studied and responded to big quakes in the region in 2004 and 2006, but its work extends to many mountainous regions at high risk of natural disasters, including in Nepal in the aftermath of its 2015 quake.

But Niigata has global ambitions as well, including in its other major centre, the Brain Research Institute. Founded in 1967, it remains the only neuroscience research institute at a Japanese national university.

Its agriculture department also maintains a collaboration with Thailand to study and develop specialized varieties of rice, including low-protein rice intended to benefit those with kidney disease.

CHUBU UNIVERSITY

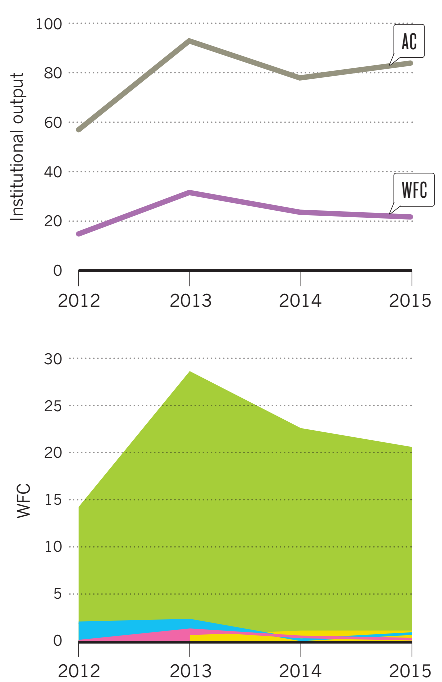
2015 WFC: 7 **2015 AC:** 19

Located in Kasugai, a suburb northeast of Nagoya, Chubu University was founded in 1938 as a technical and engineering university. In 2001 it opened a college of biotechnology and five years later added another for life and health sciences in 2006. Today it has more than 10,000 undergraduate students and 250 graduate students. The university's WFC increased from a low base, one, in 2012 to 7.4 by 2015.

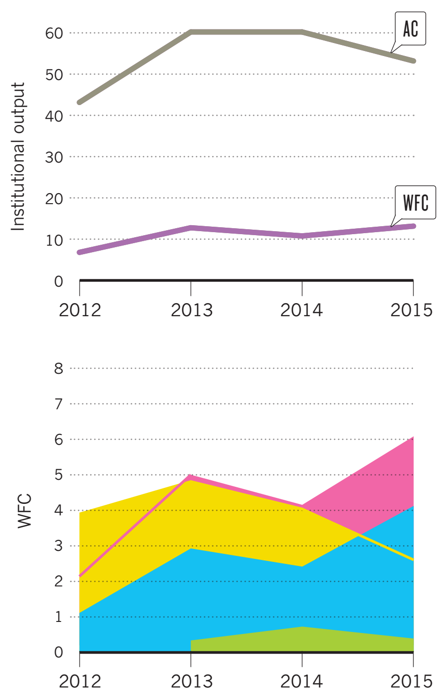
In recent years, the university has focused on recruiting top researchers including chemist Hisashi Yamamoto, who is known for his work in synthesizing organic compounds and moved from the University of Chicago to head Chubu's Molecular Catalyst Research Center.

One of the biggest projects on campus is Satoru Yamaguchi's work on superconducting power lines, which could deliver electricity through underground pipelines much more efficiently than overhead power lines. He is currently collaborating with two Japanese power companies and an internet cloud provider on a test site on the northern island of Hokkaido, a haven for wind and solar power. In 2015, they powered a data centre from solar power using a 500-m superconducting cable, one of the world's longest tests yet. By 2018, they hope to test a 2-km line that can connect to the commercial power grid. ■

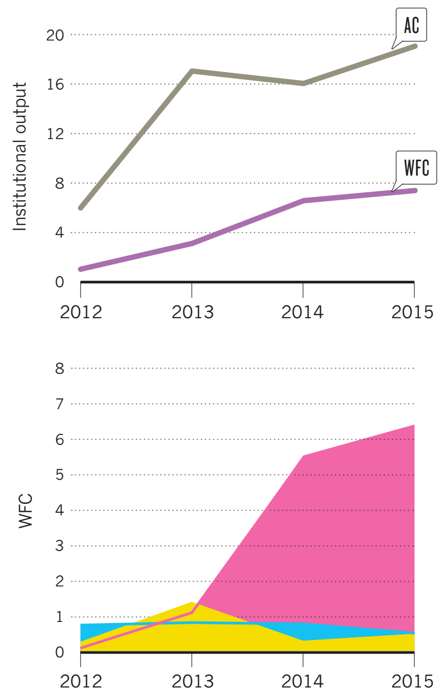
SPOTLIGHT ON LIFE SCIENCES



WELL MIXED



GROWTH SPURT



● Chemistry ● Earth & environmental sciences ● Life sciences ● Physical sciences

The Subaru telescope in Hawaii, managed by NINS, was integral in the discovery of five new planets and is a fine example of the benefits of international collaboration.

JULIE THURSTON PHOTOGRAPH /GETTY IMAGES

TOP TEAMS TO BE RECKONED WITH

Japan's institututions are increasingly joining huge collaborative research efforts, realizing that, on the big questions facing science, a problem shared is a problem halved.

BY SMRITI MALLAPATY, SIMON PLEASANTS AND NICKY PHILLIPS

These days, large-scale international collaborations are a key engine of cutting-edge science, and the past 12 months have provided some prime examples. While a single paper in *Physical Review Letters* credited more than 5,000 researchers, breaking the record for the most authors on a single article, the 2016 Breakthrough Prize in Fundamental Physics was shared between 1,377 scientists investigating neutrino oscillations.

The Nature Index is well poised to evaluate this important phenomenon because it counts the number of authors from different institutions who contribute to every paper included in the index. This is used to assign a collaboration score to institutions and countries. In Japan, the data confirms the link between quality science and collaborations. Four of the top five collaborating institutions in 2015 are also among the most prestigious — the University of Tokyo, RIKEN and the universities of Kyoto and Osaka. Rounding out the top five is the National Institutes of Natural Sciences (NINS), established by the government in 2004 to manage expensive public facilities. In 2014, almost 9,000 researchers from more than 50 countries used the facilities at NINS to conduct research. "This model of designating special institutes to care for large-scale and cutting-edge facilities shared by researchers is unique to Japan," says Amane Koizumi, a professor at NINS. Similar institutes are set up for large information systems and high-energy particle accelerators.

The prominence of NINS reflects how necessity is driving Japanese research institutions to seek collaboration. The 2016 budget for science and engineering is set at 3.45 trillion yen, down 6.4% from its peak in 2012. To help institutions cope, the government has introduced policies and programmemes to foster collaboration and resource sharing among research institutes and private companies.

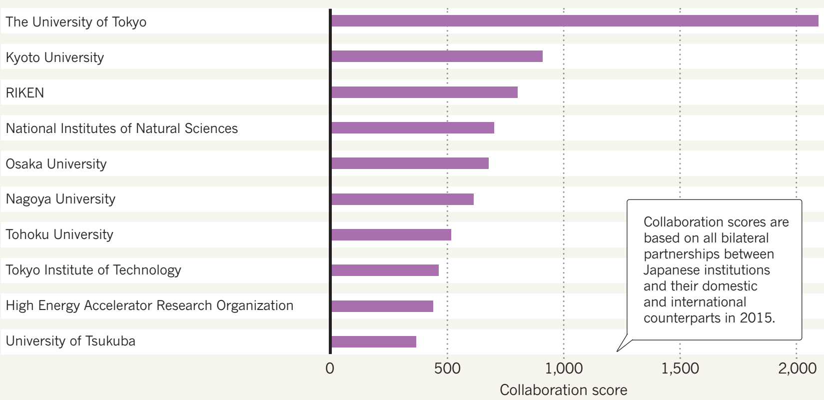
Yasuo Miake, a vice president and executive director for research at the University of Tsukuba, says another factor driving the collaboration effort in Japan institutions is government pressure to contribute to social innovation. There has been a mindset among some Japanese professors that universities should conduct science for science's sake and remain independent of government and industry, says Miake. "That mentality is changing now."

A government strategy released in 2013 identified collaboration between industry, academia and government as one of six essential principles for science policy, setting targets to double, by 2030, the number of joint projects larger than 10 million yen and projects spanning more than three years. Among the institutions whose partnership with industry lead to high-impact research, the National Institute of Advanced Industrial Science and Technology (AIST) received the index's second-highest collaboration score for its partnership with industry in Japan in 2015, following the University of Tokyo.

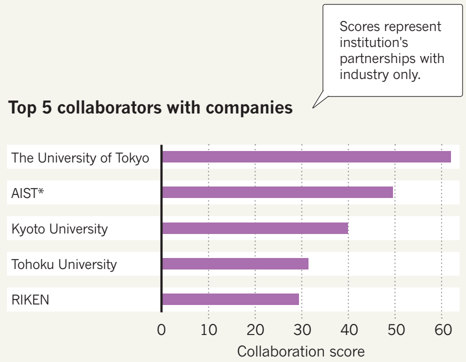
COLLABORATION CHAMPIONS

Japan's institutions are not just reaching out to domestic partners, but also internationally. The country has a wide collaboration network of academic and corporate connections across the world.

Top 10 overall collaborators



Top 5 collaborators with companies



75% The contribution of the top 10 collaborators to Japan's total collaboration score.

*National Institute of Advanced Industrial Science & Technology (AIST)

RIKEN

2015 COLLABORATION SCORE: 801

RIKEN is one of Japan's leading research institutions with roots in basic research, and wants to broaden its connection with industry. It is third on the list of Japanese institutions with the highest collaboration scores in 2015.

"RIKEN is a pure research organization, and one of our strengths is managing large projects and large facilities. These endeavours require collaboration," says a RIKEN executive director, Shigeo Koyasu.

RIKEN's focus on basic science has led to strong partnerships with other domestic and international universities with strengths in this type of research. In the index, RIKEN's top two domestic collaborators are the basic research powerhouses, University of Tokyo and Kyoto University, with which it works closely on stem-cell science.

> *"One of our strengths is managing large projects and large facilities. These endeavours require collaboration."*

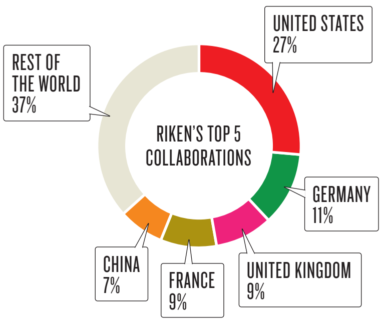
RIKEN's top international collaborators in the index are MIT in the United States and the Max Planck Society in Germany, but the institution also has major partnerships with universities in China, Korea, India and Malaysia.

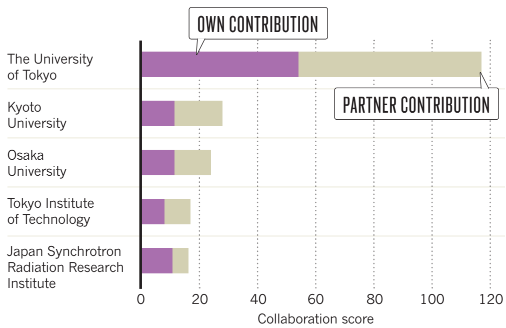
Despite having roots in fundamental research, RIKEN is moving towards solidifying partnerships with industry, says Akihiro Fujita, the head of RIKEN's Cluster for Industry Partnerships. "People in industry have not been sure how to approach RIKEN precisely because our work is in basic research, and this uncertainty may have tended to hinder progress in establishing joint research projects with industry." The Cluster for Industry Partnerships aims to bridge that divide by encouraging partnerships with the private sector.

One member of this cluster, the RIKEN Innovation Center, supports researchers to work with private companies to transfer their science into commercial products and services that solve a specific technological challenge. In addition, RIKEN's five collaboration centres with Olympus, Toyota, Rigaku Corporation, Takeda Pharmaceutical and JEOL are expected to lead to new products from basic research findings.

RIKEN'S PARTNERSHIPS

RIKEN's top 5 international collaboration hotspots; and RIKEN's top 5 collaborating institutions.





The aim of personnel at the RIKEN Innovation Center (RInC), is to support the transfer of RIKEN's many scientific breakthroughs into commercial reality through partnerships with private companies.

RIKEN

UNIVERSITY OF TSUKUBA

2015 COLLABORATION SCORE: 366

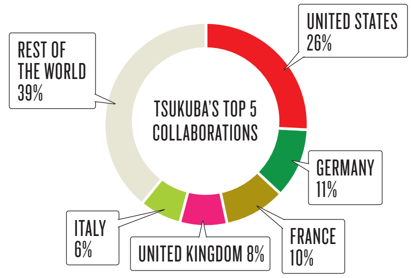
An hour by train north of Tokyo, the University of Tsukuba has taken a structural approach to fostering collaboration. Among the newer national universities, Tsukuba has broken away from some traditional university practices that would have discouraged collaboration. In the past, academic faculties at some universities were organized into large hierarchical research units around a professor. At Tsukuba, groups are much smaller and there are no restrictions on younger assistant or associate professors becoming group leaders, or principal investigators. "Because each research group is small, PIs have to look for outside collaborators to carry out their research," explains Yasuo Miake, a vice president and executive director for research at the University of Tsukuba. "Big and traditional universities have tremendous barriers between departments. Such barriers are very small at the University of Tsukuba, which allows researchers to talk more easily across departments," he says. While the university has achieved a short-term goal of doubling its income from industry through expanded partnerships with companies such as Toyota, Miake says there is more to achieve. Tsukuba is working to strengthen ties with the almost 100 industrial, academic and governmental organizations clustered at nearby Tsukuba Science City, some of which, like the National Institute of Advanced Industrial Science Technology (AIST), are already among Tsukuba's top domestic collaborators in the index.

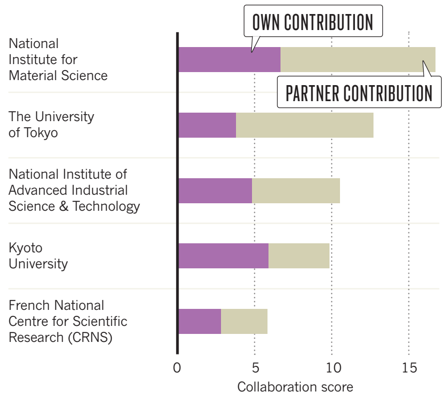
UNIVERSITY OF TSUKUBA

Synthetic organic chemists at Tsukuba University are among the small research groups at the institution.

UNIVERSITY OF TSUKUBA'S PARTNERSHIPS

University of Tsukuba's top 5 international collaboration hotspots; and top 5 collaborating institutions.





NATIONAL INSTITUTES OF NATURAL SCIENCES

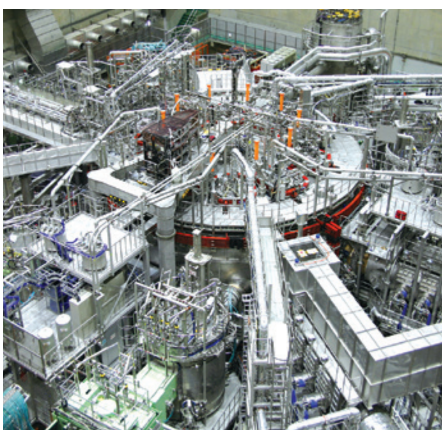
2015 COLLABORATION SCORE: 701

It took a team of 16 researchers from seven institutions, using three telescopes on two islands to discover five new Jupiter-like planets orbiting three large stars. The findings, published in *The Astrophysical Journal*, resulted from a 16-year collaboration between the National Astronomical Observatory of Japan, part of the National Institutes of Natural Sciences (NINS), and colleagues in Japan and the United States. Lead author Hiroki Harakawa from NAOJ says in radial-velocity planet searches, collaborating with researchers using other facilities adds to the number of observations required to confirm planetary systems among hundreds of targets. In this case, the observations were made using a telescope in Japan and two telescopes in Hawaii, one which belongs to the United States and the other to Japan, and managed by NINS. Among the many big research facilities it operates, the optical–infrared Subaru telescope on top of a volcano in Hawaii is the jewel in its crown.

NINS's international research alliances may continue to strengthen in the coming decade if the proposed construction of a new 30-metre telescope on the same peak in Hawaii goes ahead. A joint project by institutes in India, the United States, China, Canada and Japan, it would be the world's largest optical–infrared telescope .

In 2015, NINS researchers teamed up with scientists at 490 international and 105 domestic institutions to publish papers featured in the Nature Index, a decline from 2014. Their main domestic collaborator was the University of Tokyo, while the Max Planck Society was their most successful international partnership.

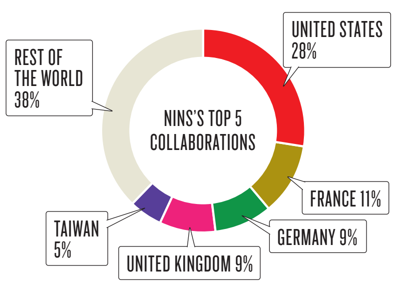
NATIONAL INSTITUTE FOR FUSION SCIENCE

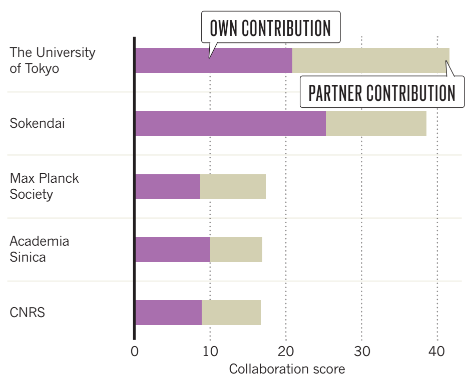


NIFS' Large Helical Device for fusion plasma research

NINS'S PARTNERSHIPS

NINS's top 5 international collaboration hotspots; and top 5 collaborating institutions.





HIGH ENERGY ACCELERATOR RESEARCH ORGANIZATION (KEK)

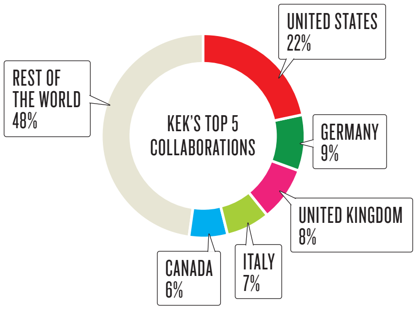
2015 COLLABORATION SCORE: 439

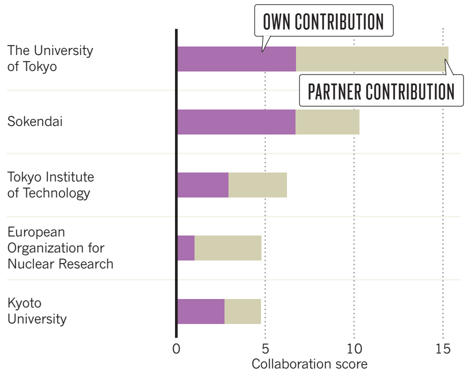
STEVE JENNINGS/STRINGER/GETTY IMAGES

Koichiro Nishikawa (left) collects his prize.

KEK'S PARTNERSHIPS

KEK'S top 5 international collaboration hotspots; and top 5 collaborating institutions.





When KEK professor, Koichiro Nishikawa, accepted this year's Breakthrough Prize in Fundamental Physics, he was doing so on behalf of one of five teams awarded for their neutrino experiments that spanned several decades.

Nishikawa is a member of the T2K particle physics experiment, a successor to the K2K programme, a long-baseline neutrino oscillation experiment involving collaborations between several countries including Japan, Germany, Italy, South Korea, Spain, the United States and the United Kingdom. These countries represent seven of the top ten nations with whom Japan partnered to produce high-quality science included in the Nature Index in 2015.

Another huge and successful collaboration involving KEK scientists is the ATLAS experiment, which observed a particle that behaved exactly like the illusive Higgs boson in 2012.

About 200 Japanese scientists, including 14 from KEK, were involved in major elements of the construction of the ALTAS detector, including the silicon strip detector to detect charged particles, the superconducting solenoid magnet and the electronics for the muon trigger chamber. "To build the large accelerator and the large detector, international collaboration is essential in terms of both budget and human power," says Kazunori Hanagaki, professor of the Institute of Particle and Nuclear Studies (IPNS) at KEK.

In the Nature Index, KEK's focus on high-energy physics reflects the main institutions with which its researchers collaborate. KEK scientists co-authored the most high-impact articles with counterparts at the INFN National Laboratory in Italy, publishing 141 papers in the 68 high-impact journals featured in the Nature Index.

NATIONAL INSTITUTE OF ADVANCED SCIENCE AND TECHNOLOGY

2015 COLLABORATION SCORE WITH INDUSTRY: 50

AIST

It's notoriously difficult to make a material that reflects light and is also thermally insulating. In August last year a collaboration between researchers from the National Institute of Advanced Industrial Science and Technology (AIST) and Kawaken Fine Chemicals Co. Ltd succeeded where others had failed.

Taking inspiration from the layered arrangement of fish scales, the research team devised a simple method for producing an alumina film that strongly reflects light and has a high thermal insulation comparable to that of wool. The method was reported in *Advanced Materials*.

The study demonstrates the importance of fostering collaborations across the academia–industry divide says Kenji Hata, the director of the CNT-Application Research Center at AIST.

"Ten or twenty years ago, Japanese companies were very strong and would hire their own researchers, but they can't afford to do that anymore."

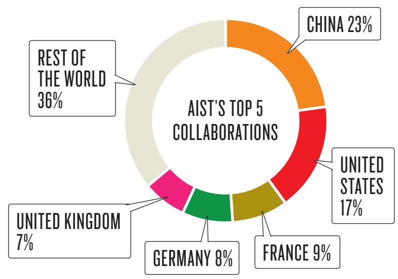
AIST was established in 2001 with a strong emphasis on applied science and collaborations with industry. The institution underwent a major reorganization in 2015 to better align its projects to meet the needs of industry, including creating a new organization in charge of technology marketing. Today, in addition to its own 2,300 researchers, about 1,800 scientists and engineers from private industry carry out research at AIST each year.

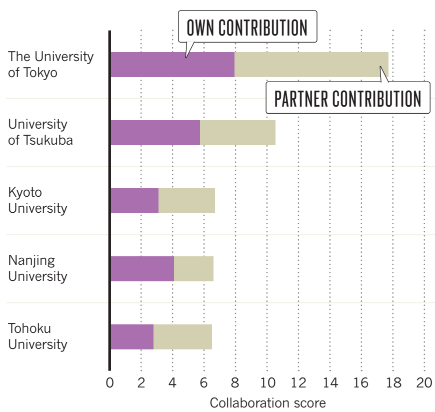
AIST campus in Tsukuba

In 2015, AIST's most fruitful industry partnership in the Nature Index was microscope manufacturer JEOL Ltd, followed by telephone communications company, NTT Group. ■

AIST'S PARTNERSHIPS

AIST'S top 5 international collaboration hotspots; and top 5 collaborating institutions.





A GUIDE TO THE NATURE INDEX

A description of the terminology and methodology used in this supplement, and a guide to the functionality available free online at natureindex.com.

The Nature Index is a database of author affiliations and institutional relationships. The index tracks contributions to articles published in a group of highly selective science journals, chosen by an independent group of active researchers.

The Nature Index provides absolute counts of publication productivity at the institutional and national level and, as such, is one indicator of global high-quality research output.

Data in the Nature Index are updated monthly, with the most recent 12 months of data made available under a Creative Commons licence at natureindex.com.

The database is compiled by Nature Publishing Group (NPG) in collaboration with Digital Science.

The list of journals tracked by the Nature Index is under review, and from 2016 will be extended to include the clinical sciences.

NATURENDEX.COM

A global indicator of high-quality research

natureindex.com users can search for specific institutions or countries and generate their own reports, ordered by article count (AC), fractional count (FC) or weighted fractional count (WFC).

Each query will return a profile page that lists the country or institution's recent research outputs, from which it is possible to drill down for more information. For example, articles can be displayed by journal, and then by article title. As in the supplement, research outputs are organized by subject area. The profile page also lists the institution or country's top collaborators, as well as its relationship with other research organizations.

nature INDEX

Home | Institution outputs | Country outputs | Customer support | FAQ

Home / Institution outputs / Institution name

Institution name

Country

Research | Collaboration | Relationships

1 January 2014 - 31 December 2014

Region: Global
Subject/journal group: All

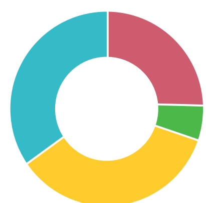
The table to the right includes counts of all research outputs for Institution name published between 1 January 2014 - 31 December 2014 which are tracked by the Nature Index.

Below, the same research outputs are grouped by subject. Click on the subject to drill-down into a list of articles organized by journal, and then by title.

Note: Articles may be assigned to more than one subject area.

AC	FC	WFC
1221	598.04	558.30

Outputs by subject



Subject	AC	FC	WFC
Chemistry	276	179.1	179.11
Earth & Environmental Sciences	95	42.73	42.73
Life Sciences	439	231.50	231.50
Physical Sciences	652	284.48	244.74

◂ Return to institution outputs

NATURE INDEX METRICS

There are four measures provided by the Nature Index to track affiliation data. The simplest is the ***article count (AC)***. A country or institution is given an AC of 1 for each article that has at least one author from that country or institution. This is the case whether an article has one or a hundred authors, and it means that the same article can contribute to the AC of multiple countries or institutions.

To get a sense of a country or institution's contribution to an article, and to remove the possibility of counting articles more than once, the Nature Index uses the fractional count (FC), which takes into account the relative contribution of each author to an article. The total FC available per paper is 1, which is shared between all authors under the assumption that each contributed equally. For instance, a paper with 10 authors means that each author receives an FC of 0.1. For authors who have joint affiliations, the individual FC is then split equally between each affiliation.

The third measure used is the weighted fractional count (WFC), which applies a weighting to the FC to adjust for the over-representation of papers in astronomy and astrophysics. The four journals in these disciplines publish about 50% of all papers in international journals in this field — approximately five times the equivalent percentage for other fields. Therefore, although the data for astronomy and astrophysics are compiled in the same way as for all other disciplines, articles from these journals are assigned one-fifth the weight of other articles (i.e., the FC is multiplied by 0.2 to derive the WFC).

The total FC or WFC for an institution is calculated by summing the FC or WFC for individual authors. The fourth measure is the collaboration score (see The Supplement).

The process is similar for countries, although complicated by the fact that some institutions have overseas labs that will be counted towards their host country totals. What's more, there is great variability in the way authors present their affiliations. Every effort is made to count affiliations consistently, with a background of reasonable assumptions.

For more information on how the affiliation information is processed and counted, please see the FAQ section at natureindex.com.

THE SUPPLEMENT

Nature Index 2016 Japan is based on data from the Nature Index, covering articles published during four consecutive years from 1 January 2012 to 31 December 2015.

Most analyses within the supplement use WFC as the primary metric, as it provides a more even basis for comparison across multiple disciplines, and in determining the relative contribution of each city or institution. Some sections and graphics also refer to collaboration score. This is a relatively new metric that is derived by adding the FC for all the bilateral relationships for that institution or country. If institution A has relationships with two others, B and C, then the collaboration score is the sum of FC for A + B and A + C. ■

KOBE UNIVERSITY

Inspiring a new generation of innovators

With the opening of its new Graduate School of Science, Technology and Innovation in early 2016, Kobe University will bring together ideas from across the sciences and commerce to forge new areas of innovation.

Kobe city, located just 30 kilometres west of Osaka, is a major historical centre of commerce and one of the first Japanese port cities to open up to the rest of the world after the end of the country's policy of isolation in the mid-1800s. This long commercial and ideological interaction with the West has given Kobe a reputation of being an international and cosmopolitan city — an important player in the modernization of Japan and a place where great things happen. And it was in this melting pot in 1902 that Kobe University has its origins.

"Kobe University started as a higher commercial school," says President Hiroshi Takeda. "This commercial tradition formed the groundwork for some of Japan's top achievements in education and research, particularly in business administration and economics, and we have produced many of the business leaders who helped lay the foundations of modern Japan."

At the same time, Kobe University has built an international reputation for education and research in the natural sciences, life sciences and medicine, from bioproduction to advanced membrane technology, infectious disease research, synthetic biology and planetology. One of Japan's greatest contributions to the life sciences — the discovery and development of induced pluripotent stem (iPS) cells — was made by Kobe University's Shinya Yamanaka, who received the 2012 Nobel Prize in Physiology or Medicine for the achievement.

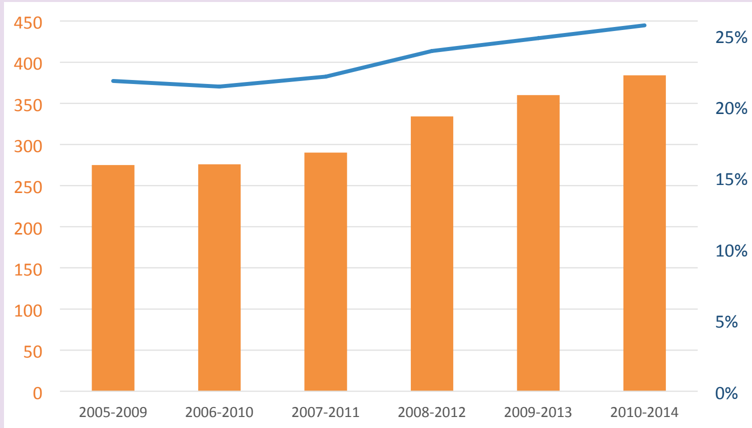
"While continuing to work towards more in-depth research in each field, we also emphasize combinations of and collaborations between different academic areas by engaging in various advanced and interdisciplinary research projects," says Takeda. "In 2007, we created a unique interdisciplinary organization, the Organization of Advanced Science and Technology, to streamline graduate school education and promote advanced research and collaboration across five of our science graduate schools. The Graduate

President Hiroshi Takeda

School of Science, Technology and Innovation will further support advanced interdisciplinary research in the natural sciences as well as equip our students with the entrepreneurial skills needed to commercialize academic research results."

Bridging the gap

The new Graduate School of Science, Technology and Innovation is unique in Japan in embedding education and capacity-building in innovation, commercialization, entrepreneurship and enterprise into a graduate science programme. With this initiative, Kobe University is aiming to nurture a new generation of science professionals with a commercial mindset and



Source: InCites™

International collaboration

Both the number (orange bars) and proportion (blue line) of internationally co-authored papers are increasing at Kobe University. Researchers at the university are actively engaging in international collaborations across a wide range of fields from astrophysics to economics.

the ability to take an idea from research to the securing of intellectual property rights, development of production technology and eventually market development.

"The so-called valley of death, where research and development does not connect to commercial success, has become an issue for industry worldwide," says Takeda. "I hope that the people we train in this new graduate school will have the ability to solve these industry problems and lead innovation. I believe that this initiative could revitalize industry in Japan, and maybe even internationally."

In yet another first for a Japanese university, the establishment of the new graduate school is accompanied by the creation of the university's own venture company, which will incubate and commercialize the research results coming out of Kobe University.

"This new Graduate School of Science, Technology and Innovation is a trial in Japan," says Akihiko Kondo, who will assume deanship of the new graduate school in April 2016. "With this graduate school, we will draw on Kobe University's strengths in bioproduction, membrane technology, communication technology and medical science, and combine them with each other, with partners from industry, and with education and training in entrepreneurship, finance and innovation to create new commercial fields of technology."

Kobe University has been involved in developing several biological production processes that are now used in industry, bringing together fields such as synthetic biology, metabolic engineering and biopharmaceuticals. It is Japan's centre for synthetic biology research — the engineering of 'designer' biomolecules and organisms — and has developed strong synergetic relationships with local biomedical companies in the medical application of synthetic biology technologies. The university also hosts the Integrated Biorefinery Center, a commercial pharmaceutical production facility and Japan's sole collaborative facility for bioproduction.

Membrane technology has long been an area of strength for Kobe University. The Center for Membrane and Film Technology is the only research centre of its kind in Japan and serves as a central research hub for advanced membrane research. "Our researchers are leaders in the recovery and separation of carbon dioxide, desalination, gas separation, membrane antifouling technology and the emerging area of forward osmosis for separating and concentrating target molecules," says Kondo.

In medical science, the Graduate School of Science, Technology and Innovation will pursue many promising areas of research, building on Kobe University's pedigree in iPS technology. "In addition to further developing iPS technology itself, we are involved in cancer therapy and drug discovery, as well as advanced vaccine development and the computationally assisted search for new drug candidates," says Kondo. "Computational science is critical to many areas of advanced research, from calculating new materials for photosynthesis and catalysis, to the design of enzymes and cells. Multiscale supercomputer simulations will open many new interdisciplinary opportunities for innovation."

A national innovation hub

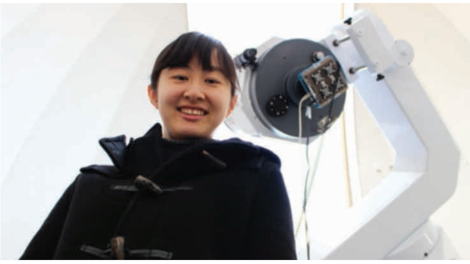
The holistic integration of research, development and commerce at Kobe University does not end at the campus gates. The university's collaboration and joint research activity extends well into Kobe city and beyond — an area awash with national research institutes and world-class research facilities, including the Port Island super science cluster, the SPring-8 high-energy particle accelerator and SACLA X-ray free electron laser, and the K computer — one of the fastest supercomputers in the world.

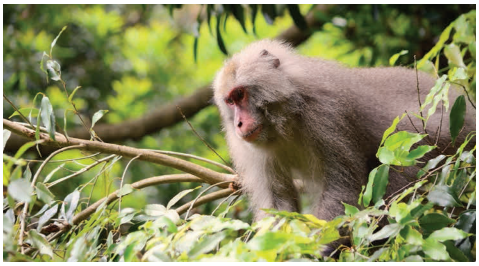
"Kobe University successfully blends tradition with innovation, and humanities with the social and natural sciences," says Takeda. "We provide special education programmes, including some offered entirely in English, with the aim of producing globally minded students. And we pursue innovation through interdisciplinary research. In this way, we will continue to strive to meet the challenge of creating new values in order to solve current issues and support the society of the future."

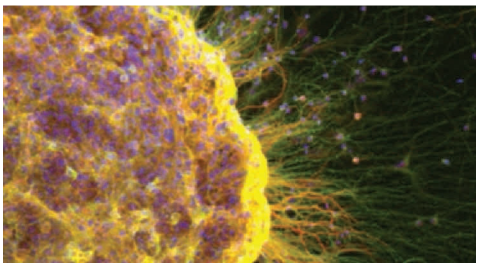
Contact

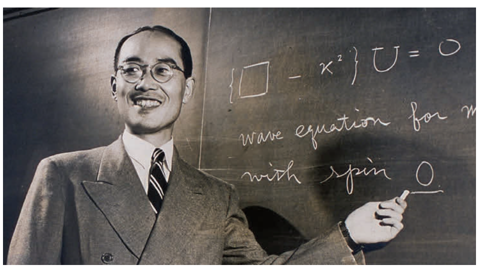
Kobe University
International Affairs Planning Division
Tel: +81 (0) 78-803-5046
E-mail: intl-relations@office.kobe-u.ac.jp

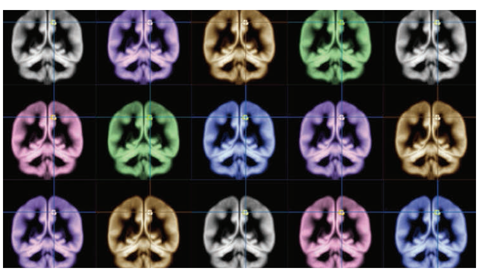
KOBE UNIVERSITY











KYOTO UNIVERSITY

To wander with purpose

A tradition of unorthodox thinking has defined Kyoto University over its 120-year history. The freedom to explore unconventional ideas allows cutting-edge science to flourish at the institution.

President Juichi Yamagiwa's ambition is to keep Kyoto University a safe haven for free thinking: free from societal pressures, free from prejudice and free from existing paradigms.

"In present-day Japan, people tend to be treated harshly when they think outside the box," Yamagiwa observes. "Universities should protect students and researchers from that. Free thinking arising from the liberal atmosphere here helps pioneer the future."

Many come to Kyoto for the opportunity to conceive and explore radical ideas. They include physicist Hideki Yukawa (second photo from bottom), the first Japanese Nobel laureate, and Kitaro Nishida, who founded the Kyoto school of philosophy. Yamagiwa is also a pioneer, spending years in the African jungle living with gorillas and studying human evolution.

Ultimately, Yamagiwa plans to make Kyoto an incubator for 'frontier science' — progressive research based on interdisciplinary communication. "Frontier science should transcend national borders and also the differences between the arts and the sciences," he says, reaffirming his intention to spread Kyoto's distinctive approach to the rest of the world. "Our researchers don't just follow the cutting edge. They are the pioneers that plough the field for others to follow," he emphasizes.

Animals and technology

And plowing through the field of great ape behavior is Fumihiro Kano, who employs the latest technologies, including eye tracking and thermal sensing, to his research. In a recent study, Kano used eye-tracking technology to examine how well apes recall one-off events stored in their long-term memories. Kano's team recorded where the apes looked while viewing video clips starring a colleague dressed as a 'bad' gorilla. The results demonstrated that apes can remember events they had seen or experienced only once. "Studying animal behavior is no longer just about direct observation and long-term training," Kano explains.

Meanwhile, Kano's colleague Masaki Tomonaga has been testing the visual perception of ponies using touch screens. In his study, ponies discriminate between different shapes by nuzzling a touch screen. Comparison of these results with those of studies involving other mammals reveals that ponies see the world as humans do, he says. "New technology is developing quickly. It opens up so many options."

Peering into germ-cell development

Behavioral scientists are not the only ones finding new uses for the latest advances. Kotaro Sasaki is breaking new ground with induced pluripotent stem (iPS) cells, whose discovery led to Kyoto University's Shinya Yamanaka being awarded the 2012 Nobel Prize in Physiology or Medicine.

Pluripotent stem cells have been used to generate just about every cell type in the body except germ cells, which include sperm and ova. Unlike other stem-cell-based therapies, which affect only the patient, those that induce germ cells could also benefit the patient's offspring.

Germ cells are difficult to study because of the ethical and technical issues involved in experimentation. "When I read about his

work, I knew I had to come back to Japan," says Sasaki, referring to the work of lab leader Mitinori Saitou. Sasaki, who was a renowned pathologist in the United States, returned to join Saitou's team in Kyoto.

Sasaki spent a month at the Center for iPS Research and Application learning how to work with human iPS cells and about the developmental process of human germ cells. In addition to illustrating key transcription interactions and signaling events, Sasaki gained insight into how epigenetic marks — traits that are inherited without changes to the DNA sequence — are 'erased' at the beginning of germ-cell development.

The team's model, still in its early stages, may form a foundation for continuing studies on germ cell lineage. "By further reconstituting human germ-cell development in vitro," explains Saitou, "we may be able to elucidate the mechanisms throughout the developmental process, from embryo to adult."

'Seeing' black holes with visible light

In the field of astronomy, Mariko Kimura (second photo from top on facing page), a master's student, took full advantage of Kyoto University's international network when she witnessed a major event unfold in front of her eyes.

Our locations:

Kyoto University maintains research and observation centres at locations throughout Japan (pictured), from Hokkaido to Kagoshima, as well as numerous offices and facilities overseas.

Kimura called affiliated astronomers when the binary black-hole system V404 Cygni generated an outburst, an event that occurs only a few times a century. Within minutes, she had international collaborators supporting her with unprecedented amounts of observational data. "We now know that we can make observations based on optical rays — visible light, in other words — and that black holes can be observed without high-spec X-ray or gamma-ray telescopes," Kimura says.

Her study indicates that the main factor triggering repetitive activity around black holes is not the quantity of mass accretion, but rather the length of orbital periods. "The study would not have come to fruition," says Daisaku Nogami, leader of the research unit, "were it not for the readily available network Kyoto astronomers have established through years of effort with a view to advancing mutual support in observation."

"Stars can only be observed after dark, and there are only so many hours each night, but by making observations from different locations around the globe we're able to gather data more comprehensively," elaborates Nogami.

Searching for where happiness happens

In Yamagiwa's haven for free thinking, other researchers delve into a question left unanswered for millennia — our own well-being.

Throughout human history, eminent scholars such as Aristotle have contemplated happiness, says Wataru Sato. He is determined to find answers with his own twist and with the help of functional magnetic resonance imaging (fMRI). "Emotions arise from our subconscious mind and spirit," says Sato. "People bond through emotions, but we can't access our subconscious minds directly. Being able to understand that part of ourselves could help us solve problems associated with the mind and spirit — including how to be happy."

At the Kokoro Research Center, scholars like Sato explore the mind and spirit, with some examining the topic from a philosophical perspective, and others, like Sato, integrating neurological insights in their studies.

President Juichi Yamagiwa

Using fMRI, Sato has identified the part of the brain that 'feels' happiness and has clarified what factors constitute happiness. According to his study, happiness is a combination of happy emotions and satisfaction with life coming together in the precuneus, a region in the medial parietal lobe that becomes active when experiencing consciousness.

For Yamagiwa, this is only the beginning. Even in the face of the uncertainty of this globalized era, Yamagiwa believes that Kyoto can lead the world with its creativity. And he is not merely talking about edgy science. "We wander in academic freedom with a purpose," Yamagiwa says. "These innovations will lead to solutions for a harmonious, sustainable, global society."

KYOTO UNIVERSITY
KYOTO JAPAN
FOUNDED 1897

Contact

www.kyoto-u.ac.jp/en
www.facebook.com/Kyoto.Univ
www.twitter.com/KyotoU_News
Tel: 075-753-2071
E-mail:
comms@mail2.adm.kyoto-u.ac.jp




RESEARCH UNIVERSITY NETWORK OF JAPAN

Japan's research universities go global

The Research University Network of Japan opens a new window to the world.

As a pioneer of 'inflation theory', Katsuhiko Sato spent much of his career grappling with questions about the origins of the Universe. Now, as president of the National Institutes of Natural Sciences (NINS), he is complementing his cosmological concerns with more earthly matters: expanding the international cooperation of Japan's research universities.

In 2013, Sato and directors from 25 research universities and institutes across Japan agreed to establish the Research University Network of Japan (RUNJ), an organization that aims to enhance international collaboration and further relationships between Japanese researchers and their counterparts on the global stage.

RUNJ was established as a part of the Ministry of Education, Culture, Sports, Science and Technology's Program for Promoting the Enhancement of Research Universities. This government initiative supports efforts to link Japanese research universities and form a network of sharing and mutual support. It includes 19 universities, together with NINS and two other members of the Inter-University Research Institute Corporation (IURIC). IURIC is a unique Japanese organization that aims to create a joint forum for researchers and other university staff with the ultimate goal of enhancing the research capabilities of academic institutions throughout the country.

"Speaking with a united voice, we can be much more effective in connecting with our global peers," Sato emphasizes. "Our goal is to act not as single entities, but as a united group of Japanese research universities, communicating our activities and recommendations not only to Japan but also around the world."

RUNJ has ambitions to engage in concrete activities, such as establishing task forces to address common issues facing universities. To aid communication in a global context, RUNJ has proposed the creation of a Japan portal for EurekAlert!, an online science news service provided by the American Association for the Advancement of Science. This highly visible communication channel would allow Japanese research universities to disseminate bilingual information to a wide global audience. RUNJ's Task Force on International Collaboration has also established Gateway, an email-based system that will expedite the circulation of information concerning research collaboration, offering a new approach to problem-solving with an eye towards international research development.

Together, these RUNJ initiatives are opening a new window to the world for Japanese researchers, allowing the forging of richer and stronger connections with their international peers.

Gateway between Japan and the world

The Research University Network of Japan (RUNJ) seeks to promote international collaboration with Japan's research universities. It seeks to provide opportunities for good networking.

contact@runetwork.jp

→ **Access several strategically placed administrators who interface with faculties.**

→ **Coordinate joint workshops and seminars and encourage international collaboration.**

EurekAlert! Japan Portal now acts as a portal between Japan's research universities and the world

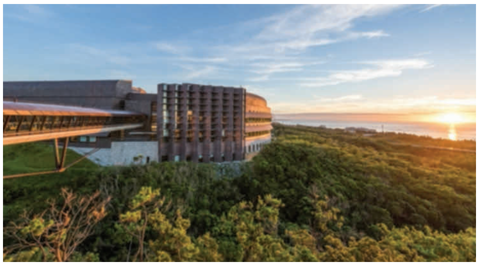
RUNJ has collaborated with the American Association for Advancement of Science to redesign their EurekAlert! and to launch a Japanese news portal to disseminate our bilingual press releases to the site's global audience.

EurekAlert! Japan portal:
www.eurekalert.org/language/japanese.php

NINS
National Institutes of Natural Sciences

Contact

E-mail: contact@runetwork.jp
Phone: +81-3-5425-1300
NINS: www.nins.jp/english/index.php
RUNJ: www.runetwork.jp/english/



THE OIST GRADUATE UNIVERSITY

A student-centred approach to education

OIST is forging a new path of graduate education — one that its students help to create

The Okinawa Institute of Science and Technology Graduate University (OIST) recognizes that many great discoveries lie at the intersections of the major sciences, where researchers make novel connections between disparate disciplines. Accordingly, it has developed a graduate programme for students to become highly proficient in their core discipline while also learning to communicate scientifically with people in completely different disciplines, such as quantum physics and cell biology.

"We want our students to achieve first-class research outcomes, to creatively address important scientific questions and to grow to their full potential as independent scientists playing leading roles in research," says Jeff Wickens, dean of the OIST Graduate School. "We encourage an international perspective and growth that is unrestricted by the traditional boundaries between disciplines."

Freedom is a vital condition of an environment that fosters creativity. At OIST, students are not admitted to specific laboratories, but are given a year to decide in which research unit they will do their thesis work. Students may start the programme with only a bachelor's degree. If they come with a master's degree, they can get credit for their additional background, but they still have a year to choose their thesis lab.

The PhD programme is individualized with a flexible curriculum that can be tailored for each student. Every student is treated as a unique individual and can select courses without the constraints of traditional departments. To help students make the most of the opportunities and develop a coherent programme of studies, they are assigned an experienced faculty member as an academic mentor to guide their course of study and assist them choose a thesis lab. Small classes, a low student-to-faculty ratio of just two to one and collaborative principles let the students shape how classes are taught. If OIST does not offer a recommended course, the student can work with a tutor in guided independent study or with visiting professors in special topics.

Instead of having traditional departmental structures, OIST has concentrations in several fields, which, while distinct, permit cross-disciplinary interactions. OIST has faculty in chemistry; physics; mathematical and computational sciences; systems biology and bioinformatics; molecular, cellular and developmental biology; neuroscience; and environmental and ecological sciences; and marine sciences. Students can take courses and conduct PhD projects in these fields and are encouraged to stretch beyond their core discipline and take some courses and laboratory rotations that are well out of their field.

From the first week, students participate in research, even before choosing a thesis lab. In each of the three terms of the first year, they conduct a research project in a new lab. Usually two rotations fall within the chosen field (theoretical and experimental physics, for example) and one far outside it (like ecology).

"Rotations are designed for the student to learn, as an insider, how to speak the language and apply the techniques used in a research unit. We challenge them to go beyond their comfort zones, to learn

research by doing research and to gain first-hand experience working alongside leading scientists and faculty members," says Wickens. "This opens up future possibilities for meaningful interactions with researchers in different fields, leading to new possibilities from the cross-fertilization of ideas and collaborative projects."

In the second year, students define their PhD thesis topics and labs, preparing a proposal that will form the basis of their end-of-year qualifying examination. Prominent international experts in the field of the student's research topic are brought to OIST to conduct an oral examination, setting a high international standard for the OIST PhD degree.

Throughout the programme, the OIST philosophy of research is promoted by offering equal access to research equipment. This encourages interaction and collaboration within and between disciplines as a way to find novel solutions and new discoveries.

The OIST PhD programme is fully funded, allowing students to progress in their studies and research without worrying about finances. They all receive an internationally competitive financial support package, comparable to those offered by other leading research universities. Most students initially live in the campus village, further strengthening the OIST community with plenty of opportunities for socializing.

OIST is truly international. Education and research are conducted in English, and the academic year starts in September. Students are encouraged to travel internationally to keep abreast of new developments, disseminate their research findings and tap into the extensive networks of OIST faculty members. This will develop future career opportunities in leading research institutes and universities worldwide.

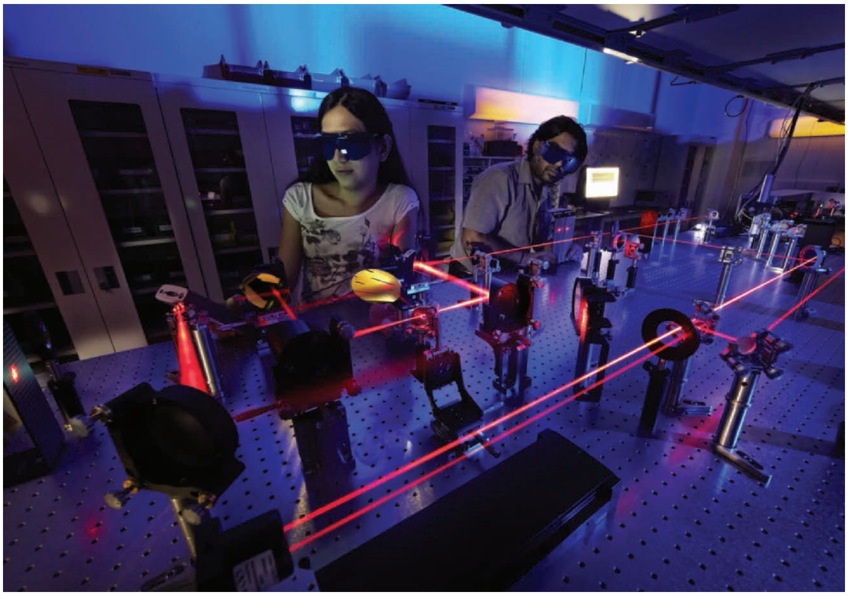
The diversity of the student population is astounding, both in terms of scientific interests and national origins. The student community is vibrant with members supported by active programmes to promote well-being and to help them support each other. To cover any gaps in academic background, self-help programmes have sprung up both spontaneously and by design, as 'skill pills' organized by the graduate school to address additional study needs as they arise. These, along with an ongoing programme of international workshops and courses, enable students to gain teaching experience.

Since competition for jobs within and outside academia is fierce, professional development opportunities are provided throughout the programme. Training is given in presenting science, research conduct and ethics, with a view to maximizing future career opportunities. Career development is seriously promoted, and a programme of seminars by student-invited speakers provides perspectives on management, leadership and entrepreneurship to help students develop into leaders who not only have outstanding research skills but also a global mindset. A constant stream of visiting speakers covers diverse topics ranging from the economics of development to how to publish in high-impact journals.

Commitment to students is one of the core values of the university. "What makes our PhD programme so exceptional is the student-centred approach," explains Wickens. "We recognize that students are individuals and design their programmes of study with them, according to their unique needs and scientific aims. We provide the resources, guidance and support they need for each step they take towards achieving their goals in research and scholarship."

Fast facts:

- → **Ratio of students to faculty members 2:1**
- → **Average class size: 4.5**
- → **Percentage of OIST students who are non-Japanese: 80%**
- → **Percentage of OIST professors from overseas: 70%**
- → **Percentage of OIST professors awarded PhDs abroad: 73%**
- → **Proportion of students in physical sciences/life sciences/ other: 45/45/10**
- → **Percentage of female students: 33%**



OIST

Contact

https://admissions.oist.jp/
https://www.facebook.com/oistedu/
e-mail: admissions@oist.jp

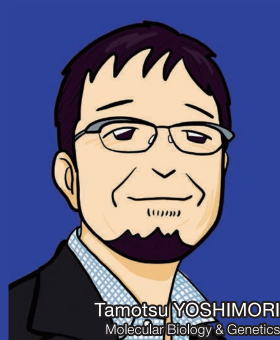
大阪大学
OSAKA UNIVERSITY
www.osaka-u.ac.jp

Research at Osaka University
ResOU
http://resou.osaka-u.ac.jp

OSAKA UNIVERSITY

Japan's Most Innovative University

18th Most Innovative in the World
By Reuters Top 100 Innovative Universities 2015



Eleven Osaka University Researchers Selected to Thomson Reuters' "The World's Most Influential Scientific Minds 2015"

"Highly Cited Researchers" are researchers selected by Thomson Reuters with theses that have a large number of citations from all over the world. In 2015, some 3,000 researchers from 21 fields were selected as performing some of the most influential researcher in the world.

Osaka University in the Global Top 40 of Nature Index (see inside)

OSAKA UNIVERSITY

Osaka University's 'open' revolution

In the lead up to its 90th anniversary, Osaka University has embraced an ambitious vision to increase openness and innovation in its research, education, community and governance in order to stimulate the kind of cross-disciplinary, cross-institutional and cross-border research needed to meet global challenges.

One feature that has characterized research and technological advance since the turn of the century has been the steady rise and remarkable fruitfulness of interdisciplinary and inter-institutional collaboration. In its bold vision for the next six years, Osaka University has its sights set on taking this type of collaborative openness to a new level, opening the laboratory doors to the broader research and industrial community and also to society at large.

"Our vision emphasizes openness," says the university's president, Shojiro Nishio. "We want to break down the barriers between departments on campus and those separating the university from the outside world. We plan to open our academic findings and research outcomes to society through our five 'pillars' of open education, open research, open innovation, open community and open governance. By interweaving knowledge from different disciplines and working together with diverse participants in society, we aim to evolve into a place where new knowledge can be co-created together."

One key initiative of the university's vision for open research is the development of a big data platform designed to centralize and share all of the accumulated data obtained through experiments and simulations in every academic discipline across the university, as well as measured in various fields of society. "Using this platform, we will be able to cross data between and among fields, including humanities and social sciences dealing with ethical, legal and economic issues, to create new integrated fields of study that can produce social and economic value," explains Nishio. "Our focus on the contact point between science, technology and society will foster interdisciplinary research that deals with global challenges. We call it open research."

Osaka University embarks on this open research revolution from a position as Japan's most innovative university and among the most innovative institutions in the world, ahead of global heavyweights like Caltech, Johns Hopkins University and the University of Cambridge, according to Reuters 2015 Top 100 Innovative Universities. Osaka University's ability to innovate through generating original research and creating useful technology with economic impact stems from its broad disciplinary spectrum and a long history of challenging convention.

With its roots in the famous Kaidokudo and Tekijuku schools that were established in 1724 and 1838 respectively and produced an extraordinary number of Japan's modern luminaries, the 'Osaka spirit' has emboldened the university community as a place where future leaders can rise above fixed social standings to engage in solving the social issues. Today, this means taking the leading global research through open dialogue and innovative thinking.

"Our university was founded with the support of civil society, and due to this tradition, we are open to society and have been actively promoting collaboration with industry for many years," explains

Nishio. "We were the first university in Japan to establish joint collaborative laboratories on campus through our Industry on Campus initiative. The university is expanding in the more dynamic University–Industry Co-Creation programme. Osaka University continues to be one of the country's leading institutions — we had the highest ratio of joint papers with industry among national comprehensive universities in Japan in 2014 and are active in local and international patent registrations."

Research with impact

As a comprehensive educational and research institution, Osaka University covers a full repertoire of academic disciplines. Among them, the university has cultivated particularly strong research teams in the fields of immunology, robotics and the emerging field of photon science and technology.

The university is a renowned international hub for immunology research and hosts the prestigious Immunology Frontier Research Center (IFReC), one of nine exclusive World Premier International research centers in Japan that have been singled out by the government for significant long-term funding. "Since the beginning of its research activities in 2008, IFReC has published more than 1,000 papers in a wide variety of immunology fields, with a large proportion in high-impact international journals," says executive vice president, Toshiya Hoshino. IFReC covers a broad range of research fields from autoimmune diseases to malaria, osteoporosis and metabolic syndrome, incorporating the latest measurement and simulation technologies. The work of the centre is aimed at devising new and more efficient development strategies for vaccines and immune therapies.

"At the IFReC, we have a number of world-renowned researchers in the field, including Shizuo Akira and Shimon Sakaguchi," says Hoshino. "Both are recipients of the Thomson Reuters Citation Laureate prize and the prestigious Canada Gairdner International Award." Akira discovered that receptors in the body's cells sense microbes and spur the immune system to combat infection and develop long-term immunity; this discovery could facilitate the development of therapies for cancer and allergies. Sakaguchi is investigating regulatory T cells as the immune cells that suppress immune reactions. His work contributes to the understanding of these enigmatic immune cells and has guided research into various physiological and pathological immune responses.

Osaka University is also globally regarded for its cutting-edge research on robotics. Hiroshi Ishiguro, for example, is a world leader in the development of human-like androids as a platform for research aimed at understanding how people converse with and operate lifelike robots. While Ishiguro's team deals with human reactions, the group of Minoru Asada focuses on the cognitive development of robots themselves. Asada's research addresses one of the most fundamental issues for the future of robotics: how cognition and the identification of self and others can be developed in robots, and how this relates to the possibility of artificial empathy. This requires research on the full developmental trajectory, crossing over with neural dynamics as the basis for understanding the interactions between the body, brain and environment that generate rich behaviours.

With more than 100 photonics laboratories, Osaka University is among the most

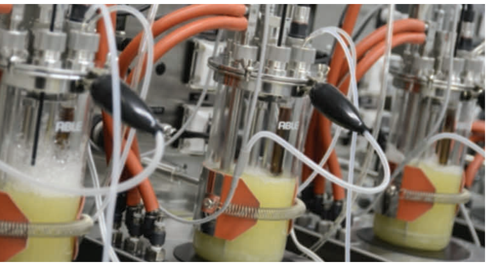
active research institutes in the world in the field of photon science and quantum beam technology. "We conduct world-class research in nanophotonics, power photonics, plasma photonics, X-ray optics and beam optics," says Ryosuke Kodama, one of the world's leading researchers in photon science — a field that promises to yield innovative technologies with exciting applications in industry and medicine. Researchers at Osaka University collaborate extensively with other national and international leaders in photon science, covering a diverse variety of fields including life science, materials science, high-energy-density science and high-energy physics, supported by the Harima Center for Photon Science located at the SPring-8 national synchrotron facility.

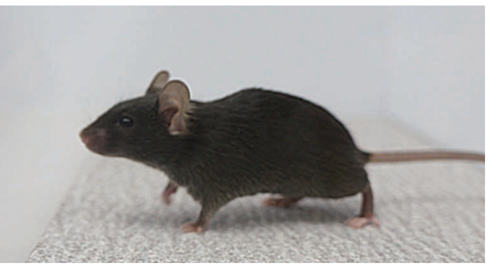
"With the distinction of being selected as a Top Global University by the Japanese government and Osaka's ranking as Japan's most innovative university," says Hoshino, "we have a responsibility to fulfil our role as a leading national university by providing rich courses and programmes, nurturing global-mindedness, and promoting research and innovation in the most advanced science and technology, humanities and social sciences to solve issues of global concern. This is our 'open' revolution."

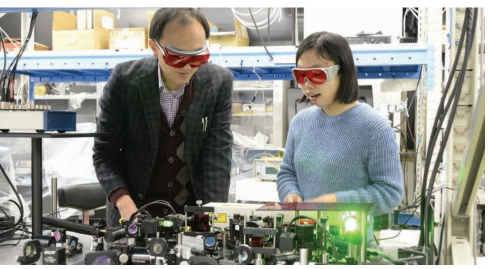
OSAKA UNIVERSITY

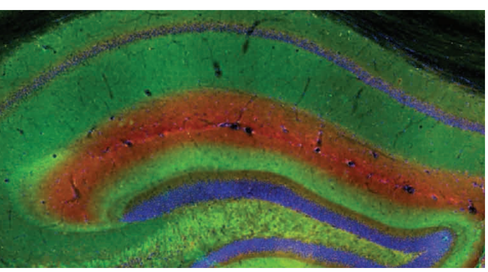
Contact

E-mail:
global-ou@ml.office.osaka-u.ac.jp
Tel: + 81 6 6105 5886
Website:
www.osaka-u.ac.jp/en/index.html
resou.osaka-u.ac.jp/en/









RIKEN NATIONAL SCIENCE INSTITUTE

Crossing disciplines, borders and scales

Founded almost a century ago in 1917, today RIKEN is Japan's largest research institution in the natural sciences. It has a prestigious history of conducting research in a wide range of specialized scientific areas. Throughout its history, RIKEN has also strived to integrate research across disciplines to deepen our knowledge of the comprehensive nature of our physical and biological environment. This vision has resulted in a better understanding of how phenomena on different scales — from the atomic level to the ecosystem and beyond — shape the world we experience.

Here, we take a closer look at two RIKEN centres at the forefront of integrated research: the Nishina Center for Accelerator-Based Science and the Center for Life Science Technologies.

Integrated nuclear physics research

On the last day of 2015, RIKEN gave a press conference (bottom photograph on the next page) to announce that just that morning, Kosuke Morita from the Nishina Centre for Accelerator-Based Science had received news from the International Union of Pure and Applied Chemistry that his group had been recognized as the discoverers of a new chemical element, element 113. It was a historic milestone — the first time an element had been discovered in Asia.

This discovery was made possible by one of the most powerful and versatile heavy ion accelerators in the world. The Nishina Centre's RI Beam Factory, where the research was conducted, is a tremendous instrument made from a linear accelerator, a series of cyclotrons, including the world's first superconducting ring cyclotron and a superconducting radioactive-isotope beam separator. The RIBF is a leading facility for accelerator-based research into the nature and origin of the nucleus and its elementary particles.

At the Nishina Center, investigators combine experimental research on particles, atomic nuclei and solid-state materials using protons, mesons, muons and other particles with theoretical methods for exploring the most fundamental laws of physics. This combined approach, together with a clear mandate to apply new elementary findings to medicine, agriculture and other natural sciences, sets the Nishina Center apart from other nuclear physics institutes.

According to Hideto En'yo, director of the centre, "The Nishina Center is also an exponent of the global reach RIKEN has gained with its research initiatives. The centre operates two facilities abroad. In the UK, we have a muon facility at the Rutherford Appleton Laboratory that provides access to an intense pulsed muon beam; with this muon beam, researchers can study the electromagnetic properties of insulating, metallic, magnetic and superconducting materials. In the USA, we have established a combined theoretical, experimental and computing group at Brookhaven National Laboratory's Relativistic Heavy Ion Collider to train young scientists in the physics of strong interactions, including spin physics, lattice quantum chromodynamics and the physics of relativistic heavy-ion collisions. Our facilities are used by scientists around the world both for basic

research and applications."

The discovery of element 113 was a first step in the centre's long-term goal of attaining one of the holy grails of nuclear physics — reaching the so-called island of stability. The island of stability theory predicts that, for elements heavier than uranium, a set of heavy isotopes will be found exhibiting a magic combination of protons and neutrons that results in increasing rather than decreasing stability.

According to En'yo, "We have already begun our search for elements 119 and 120, our next goals in the search for super-heavy elements. Other milestones on this quest include the discovery of 300 new isotopes. So far, we have detected 150."

Ultimately, the Nishina Center is on a quest to unravel the origin of the Universe by better understanding the inner workings of the atomic nucleus and thus reveal the processes by which heavy elements came to be. "The RIBF is producing a variety of data on unstable nuclei, nuclear reactions, nuclear structures, shapes, sizes, as well as many new and interesting features of exotic nuclei," notes En'yo.

Comprehensive living systems science

RIKEN's Center for Life Science Technologies (CLST) is another champion of RIKEN's focus on integrated research facilities. It was established to advance the understanding of how living systems work by integrating mechanistic insights from the atomic scale through molecular function and cellular organization to the whole-body level. This goal is achieved by approaching research from a dynamic and system-wide perspective, rather than a traditional static and reductionist angle.

"Integration in life science means orchestration of a variety of molecules in some architectonic features — CLST is mostly aiming to develop novel methodologies and techniques to follow this orchestration," explains CLST's director, Yasuyoshi Watanabe. "These technologies enable the analysis of big molecular complexes, exploration of their regulatory mechanisms and imaging the dynamics of their molecular interactions."

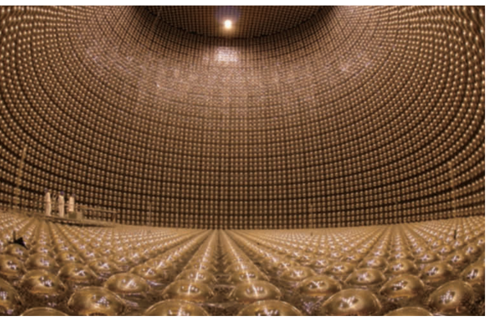
The centre's mission is driven by the close interactions among its three divisions: structural and synthetic biology, genomic technologies, and biofunction dynamics imaging. Integration of technologies and research outcomes from these three divisions allows CLST to chart a clear course towards drug discovery, new therapies and healthcare innovation.

One of the areas CLST focuses on is the characterization of disease through the integrated and multimodal analysis of structural, topological and functional data that reveal the pathophysiology and molecular phenomena underlying pre-disease and disease states. "We have achieved much in this field, not only in experimental animals, but also in humans by using a multiangle approach — total omics analyses, interaction studies of neuroimmunoendocrine systems, positron emission tomography studies, functional magnetic resonance imaging studies and magneto-encephalography studies," says Watanabe. "The ultimate goal is to find ways to prevent disease by diagnosing it and initiating treatment before its onset."

Translating new scientific and medical insights gathered at CLST into applications and therapies follows multiple paths including strong partnerships with industry, medical and academic institutions. And internally, CLST provides support for several RIKEN translational initiatives such as the RIKEN Program for Drug Discovery and Medical Technology Platforms and the RIKEN Preventive Medicine & Diagnosis Innovation Program. These collaborations also seek to improve health outcomes on a more long-term time scale. "A perfect example of this is a recently approved project spearheaded by CLST and comprising local governmental offices and more than 40 healthcare, industrial and academic collaborators," explains Watanabe. "This project aims to optimize personalized medicine approaches under the umbrella of the Compass for Healthy Life Research Complex project."

Contact

www.riken.jp/en
www.riken.jp/en/careers
www.facebook.com/RIKEN.english
www.twitter.com/riken_en
Tel: **+81 48 462 1225**
E-mail: **pr@riken.jp**



© Kamioka Observatory, ICRR, The University of Tokyo

THE UNIVERSITY OF TOKYO

Neutrinos and beyond — Opening a new era of cosmic-ray research

Takaaki Kajita, a recipient of the 2015 Nobel Prize in Physics, and other researchers at the University of Tokyo's Institute for Cosmic Ray Research (ICRR) have been exploring new realms in particle physics research. Kajita's work on neutrinos and related research at ICRR is leading the world in this field.

The path to a Nobel prize

Particle physics and astrophysics are among the most active research fields at the University of Tokyo, and its Institute for Cosmic Ray Research (ICRR) is leading the world with explorations in these areas. ICRR is best known for its research on neutrinos using the world's largest underground neutrino detector, Super-Kamiokande. The detector is located in a mine in central Japan and is filled with 50,000 tons of pure water. Neutrinos, which are electrically neutral subatomic particles, exist in three different forms — electron, muon and tau. They are generated in various places, including the centre of the Sun and the Earth's atmosphere. Although neutrinos are the most abundant particle in space after photons, they are hard to observe because they rarely interact with matter. But they hold the key to many questions about how the Universe was formed.

Researchers at ICRR have discovered many important properties of these enigmatic particles, including the fact that neutrinos oscillate, or change from one type to another, as they travel through the Earth. One implication of this finding is that neutrinos have mass — a conclusion that has forced scientists to reassess the standard model of particle physics, which predicts that neutrinos are massless. These discoveries, which resulted in Kajita's 2015 Nobel Prize in Physics, were made a team led by him in 1998.

Kajita acknowledged his success owed a lot to the strong support he received from two mentors and former supervisors — Masatoshi Koshiba and Yoji Totsuka. Koshiba was awarded the 2002 Nobel Prize in Physics for detecting neutrinos produced in supernovae using Kamiokande, the predecessor of Super-Kamiokande. Totsuka led the Super-Kamiokande project as Koshiba's successor. Totsuka's contribution was so great that many believe he would have shared the Nobel Prize with Kajita if he were alive.

Kajita's award-winning work dates back to 1986 when he earned his PhD for researching proton decay. Soon after that, Kajita began to upgrade the software he was using to separate the proton-decay signal from the background noise produced by atmospheric neutrino interactions. He discovered that his data did not match theoretical calculations and began to search for the cause of this discrepancy. He was unable to find any problems with the data analysis or the simulation of atmospheric neutrino interactions. The Kamiokande collaboration reported this result in 1988, but the academic community was sceptical. "The result disagreed with those obtained by other researchers." Kajita recalls. "But Koshiba and Totsuka understood the importance of the Kamiokande data and gave me constant support."

The breakthrough came in 1996 when Super-Kamiokande began operating; within two years Kajita's team had definitively

proved that neutrinos oscillated. They presented this finding at the International Conference on Neutrino Physics and Astrophysics. "To my surprise, the audience welcomed my presentation and gave me a standing ovation," Kajita says.

More projects with Super-Kamiokande

Since then, neutrino research has been steadily advancing. Some 120 researchers from seven countries take turns to analyse the data that is being continually generated by Super-Kamiokande. The facility is operated under the supervision of Masayuki Nakahata, head of Kamioka Observatory and solar neutrino specialist. Meanwhile, Yoichiro Suzuki, deputy director of the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU) at the University of Tokyo, greatly contributed to explaining the discrepancy between data and theory for solar neutrinos, which had long puzzled researchers. Suzuki and Kajita were awarded the 2016 Breakthrough Prize in Fundamental Physics for their roles in the discovery and study of neutrino oscillation.

Neutrino research will be further accelerated in the next few years with the upgrade of Super-Kamiokande, which will allow it to detect supernova neutrinos more efficiently. There are also plans to build another detector, Hyper-Kaimokande, whose volume will be 20 times greater than that of Super-Kamiokande. This mega-detector is expected to begin operation in 2025 and the discovery of particle–antiparticle asymmetry in neutrinos is one of the major goals for it.

Gamma rays and gravitational waves

ICRR's exploration of new research fields goes beyond neutrinos. The institute is participating in the international project to build the Cherenkov Telescope Array (CTA), a next-generation observatory of very high-energy gamma rays. CTA will detect the highest-energy photons ever observed, which will lead to a deeper the understanding of star formation in the evolution of the Universe and many other events that cannot be observed using current technologies. In the CTA project, ICRR's Cherenkov Cosmic Gamma Ray Group is developing large-scale telescopes in collaboration with researchers in Japan, the USA and Europe.

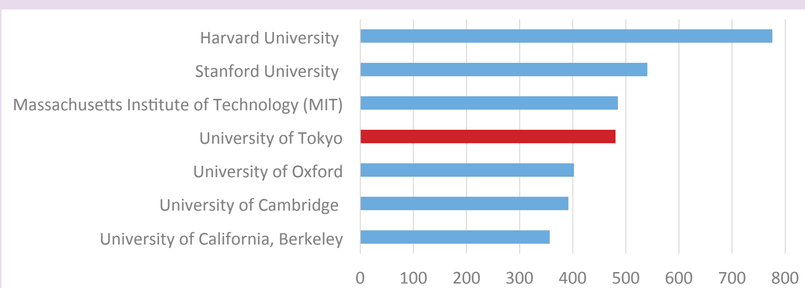
Another important project at ICRR is the Large-Scale Cryogenic Gravitational Wave Telescope (KAGRA). The first-stage facility was completed in October 2015. Gravitational waves are the curvature of space–time and they propagate at the speed of light according to Einstein's theory of relativity. On 11 February 2016, the Laser Interferometer Gravitational-Wave Observatory (LIGO) in the USA and its international collaborators made the momentous announcement that they had detected gravitational waves. "Detection of gravitational waves is very difficult, but the research — once thought impossible — is moving on thanks to the steady advance of science and technology and the endeavors of many researchers," says Shinji Miyoki, associate professor of ICRR and a member of the KAGRA project.

Researchers are now testing the laser interferometer and plan to commence its operation with cryogenic mirrors within the next year. Already, more than 200 researchers from around the world are participating in the KAGRA project.

Worldwide collaboration in this area is expected to accelerate. ICRR is poised to contribute to this new field of gravitational-wave astronomy by completing the construction of KAGRA and participating in international observation networks as soon as possible. Through collaborating with researchers around the world, ICRR aims to elucidate the mysteries of the Universe, by, for example, detecting the birth of black holes formed by the merger of binary neutron stars.

Promoting cross-border cooperation

All ICRR's projects are undergirded by the efforts of graduate students and young researchers, who perform vital activities ranging from the construction of equipment to data analysis. This is true for all the research done at the University of Tokyo. Young scientists are essential for research, and they sometimes create new research fields. As a world hub for knowledge collaboration, the University of Tokyo is supporting diverse research activities and developing an environment where young researchers can explore without constraints. To this end, the university makes proposals to the government as well as reviewing and reforming internal systems and publicizing its research activities. Support for young scientists, who will play a key role in next-generation research, is vital and should be made available by both universities and governments. As collaborative projects between researchers in different countries are becoming more prolific, it is crucial to create an environment in which researchers can flexibly and actively participate in global projects, interact with scientists around the world and create new values.



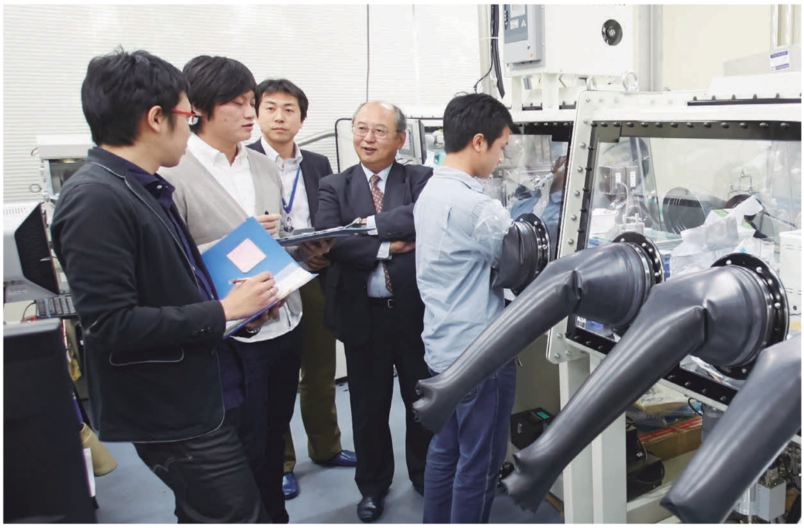
Nature Index ranking

→ **The publication output of the University of Tokyo is ranked fourth in the world in terms of the weighted fractional count (data period: 1 December 2014–30 November 2015).**

THE UNIVERSITY OF TOKYO

Contact

Website: **www.u-tokyo.ac.jp/en**
www.facebook.com/UTokyo.News.en
www.twitter.com/UTokyo_News_en
www.youtube.com/user/UtokyoPR
Phone: **+81-3-3811-3393**



TOKYO METROPOLITAN UNIVERSITY

Tokyo's strategic partner

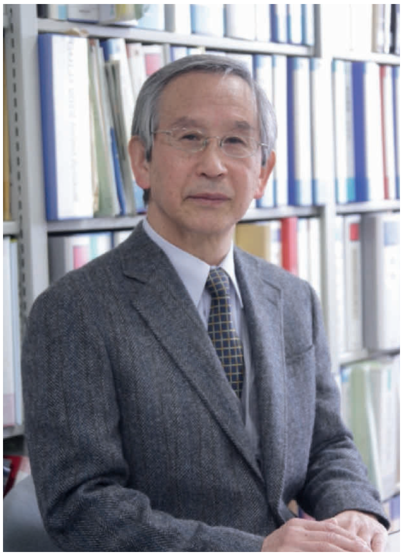
Established in 2005 but with roots dating back more than 80 years, Tokyo Metropolitan University (TMU) serves not just its students but also the metropolis of Tokyo. With the establishment of eight new world-class research centres, TMU is set to help shape the future of Tokyo and also of the world beyond.

Ranked among Japan's top universities, Tokyo Metropolitan University is highly regarded for the quality and relevance of its research. Owned by the Tokyo Metropolitan Government, TMU effectively serves as the government's think tank for solving some of the many challenges faced by the megacity of Tokyo. This unique relationship allows TMU and its researchers to contribute directly to the future of the metropolis in very tangible ways.

Pursuing 'the ideals of metropolitan society', TMU through its 28 academic departments focuses on the issues and disciplines relevant to megacities in general — such as urban infrastructure, energy and environment, health and welfare, and population ageing — as well as topics more specific to Tokyo, such as advanced disaster prevention and management.

The hosting of the 2020 Olympic and Paralympic Games by Tokyo has brought the important relationship between TMU and the Tokyo Metropolitan Government to the fore. TMU has been charged with developing some of the key technologies that promise to make the 2020 Games truly unique, engaging in research projects in areas including smart communications, advanced disaster prevention and welfare, urban environment, volunteering, regional ties and sports. The projects emphasize the practical application of research in connection with the city, governmental and municipal organizations, foreign and domestic research institutions, and business.

To stimulate and accelerate specific areas of world-class research, TMU began establishing dedicated research centres in 2014 with generous funding from the metropolitan government. The university now has 12 research centres specializing in space science; genomics and bioinformatics; artificial photosynthesis; gold chemistry; language, brain and genetics; water system engineering; community-centric systems; climatology; big social data; child and adolescent poverty; quantitative finance; and building a hydrogen energy society.



Masatake Haruta, director of the Research Center for Gold Chemistry.

Going for gold

The first of the new research centres to be established was the Research Center for Gold Chemistry. It is housed in its own new building and is under the guidance of one of TMU's most respected professors, Masatake Haruta. "In 2012, TMU hosted the Sixth International Conference on Gold Science, Technology and its Applications, and this was in many ways the catalyst for our new research centre,"

he explains, "but the history of our gold research goes back to 1982, when I first reported that gold nanoparticles can be chemically reactive."

Gold is an extremely stable metal — in its familiar form it does not oxidize or react with other chemicals, a property that underpins its value as a precious metal. Yet thanks to research pioneered by Haruta, it is now known that as nanoparticles of 2–5 nanometres in size, gold can react usefully with certain chemicals. For example, when fixed to a base metal oxide, gold nanoparticles can oxidize carbon monoxide at temperatures as low as –70 degrees Celsius. "Not only does gold display catalytic activity, it does so at lower temperatures than conventional catalytic metals such as platinum," says Haruta. "The discovery revealed a rich and unexplored chemistry within gold."

Haruta found that the electronic properties of gold vary as particles gets smaller. This means that gold nanoparticles of different sizes react differently, opening the way to design reaction-specific gold catalysts. "We are now looking at clusters of fewer than 200 atoms, where we can expect new catalytic properties," he says. "We have also found structural specificity. This is the focus of our research at the Research Center for Gold Chemistry, and our point of difference with other research groups around the world."

Already the discovery of gold nanoparticle catalysis has changed several industries. Gold is now used for vinyl chloride production and for the catalysis of methyl methacrylate — the precursor of acrylic resin. Haruta sees the most promising applications in the production of chemicals. "Gold has the potential to dramatically change organic synthesis, replacing complex and energy-intensive synthetic routes with simpler, more efficient chemistry. It also has great promise for environmental applications, like air and water purification. We are now working on improving the catalytic life of gold nanoparticles, as well as recycling and improving stability. I think gold has a very bright future in chemistry."

A hydrogen-fueled future

With the Tokyo Olympics just four years away, one of the great challenges put to the TMU by the Tokyo Metropolitan Government is to realize the government's goal of making the Tokyo Olympic Games a demonstration of a 'green energy city' based on hydrogen energy.

"Tokyo is one of the biggest cities in the world," says Kiyoshi Kanamura, director of TMU's new Research Center for Building Hydrogen Energy Society. "To realize a green energy system using hydrogen fuel in Tokyo, we need to research and develop not only hydrogen fuel cells, but also the necessary infrastructure, such as a network of hydrogen refueling stations, and hydrogen production and carrier systems. Our new research centre will provide comprehensive and cross-cutting research for the realization of a hydrogen energy society across a wide range of disciplines."

The work of the hydrogen research centre, the 12th to be established under TMU's new research centre directive, coordinates closely with that of the research centres for artificial photosynthesis and gold chemistry. "We want to use natural energy to produce hydrogen, such as biomass and solar energy, in order to realize a CO_2-free hydrogen production system," says Kanamura. "The Tokyo Metropolitan Government has committed to exhibiting Japan's hydrogen energy society to the world at the 2020 Olympics, and the TMU through this research centre will strive to make that a reality. The outcomes from the centre, the fundamental results and know-how, will then be transferred to other megacities around the world."

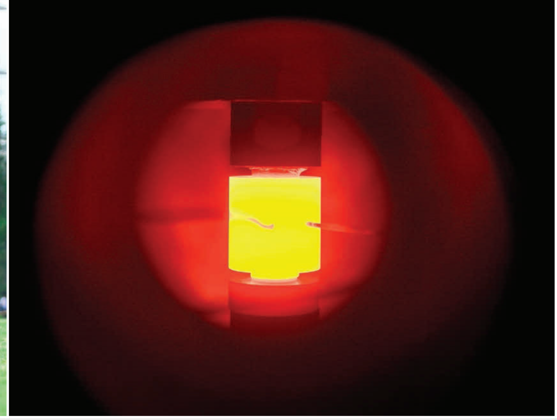
Research Centres for:

- **Space Science**
- **Genomics and Bioinformatics**
- **Artificial Photosynthesis**
- **Gold Chemistry**
- **Language, Brain and Genetics**
- **Water System Engineering**
- **Community-Centric Systems**
- **Climatology**
- **Social Big Data**
- **Child and Adolescent Poverty**
- **Quantitative Finance**
- **Building Hydrogen Energy Society**

TMU
TOKYO METROPOLITAN UNIVERSITY
& TOKYO

Contact

Website:
www.tmu.ac.jp/english/index.html
Phone: +81-42-677-2728
Fax: +81-42-677-5640
E-mail: ragroup@jmj.tmu.ac.jp



Tokyo University of Science

Building a better future with science

Strong ethics and an emphasis on multidisciplinary research underpin efforts at Tokyo University of Science to solve global challenges

Since its establishment in 1881, the Tokyo University of Science (TUS) has been progressing science and technology at a dizzying pace. With the emergence of environmental and energy challenges as critical global issues, the university's founding principle of advancing science and technology in harmony with nature is more relevant than ever before. Within a strong ethical framework, scientists and engineers at TUS are striving to solve global challenges and make the world a better place through science.

Undergraduate programmes at TUS provide students with a solid foundation in specific disciplines for development and success, while graduate students are encouraged to hone their skills in an interdisciplinary environment. In particular, students in the advanced stages of their studies can conduct research at the frontiers of their disciplines, while interacting with accomplished scholars and scientists. Research and education go hand in hand at TUS, where researchers tackle important problems that often require scientific knowledge from multiple disciplines.

TUS has established various institutes, including the Research Institute for Science and Technology and the Research Institute for Biomedical Sciences, to actively pursue research that pushes disciplinary boundaries both within the university and globally. It is committed to achieving these research goals with meaningful interdisciplinary cooperation based on a thorough understanding of the foundations of academic fields, and by transcending the distinction between theory and application. TUS hosts nearly 30 research centres organized under the auspices of these interdisciplinary institutes, such as the Center for Fire Science and Technology and the Photocatalysis International Research Center. These research centres and projects connect different fields and contribute to the development of science and technology in emerging multidisciplinary fields.

Photocatalysis is one of the university's research strengths. Akira Fujishima, the president of TUS and director of the Photocatalysis International Research Center, along with his colleagues, introduced the world to semiconductor photocatalysis on titanium dioxide during his graduate study. Since then, photocatalysis has become an important area of science and technology that may provide solutions to the ever-increasing demand for energy as well as environmental concerns. The Photocatalysis International Research Center researches self-cleaning properties, environmental purification and artificial photosynthesis. The centre has designed products and materials that are accessible to the public to increase awareness about how photocatalysis is necessary for a green and sustainable environment. The centre actively collaborates with other research groups and hires young researchers from across the world with the goal of being a top-class photocatalysis research centre.

TUS also researches fire science and supports the development of safety technology designed to protect lives and property. The Center for Fire Science and Technology is establishing a network for promoting fire safety information in Asia, with the aim of controlling fire risks in various Asian cities. It is equipped with a laboratory building whose scale and functions are world class. TUS also has Asia's first specialized graduate school in fire science, the Graduate School of Global Fire Science and Technology, which trains young researchers and engineers.

Research and development at TUS have received global acclaim. Yoichiro Iwakura was ranked by Thomson Reuters in the top 110 of Highly Cited Researchers in immunology in 2014 and 2015. He is studying the generation and analysis of human disease models using transgenic techniques.

TUS takes great pride in its broad scope and illustrious history. It remains fully committed to the mission of conducting education and research to improve the world through science and technology.

TOKYO UNIVERSITY OF SCIENCE

Contact

Tokyo University of Science
1-3 Kagurazaka, Shinjuku-ku
Tokyo 162-8601, Japan
E-mail: koho@admin.tus.ac.jp
Website: www.tus.ac.jp/en/

NATURE INDEX PROFILE
TOKYO UNIVERSITY OF SCIENCE

TUS's mission for education and research is to use science and technology for the sake of nature, humanity, society and their harmonious development.

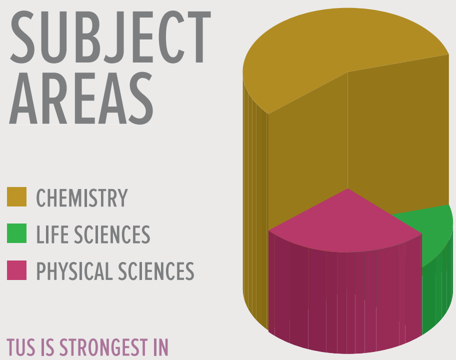
Tokyo University of Science (TUS) was founded in 1881 and is one of the oldest science and technology focused private universities in Japan. At its four campuses in the Tokyo metropolitan area and at its sister university TUS, Suwa, TUS researchers concentrate on their mission to build a better future with science.

PAPERS PUBLISHED

77

IN THE NATURE INDEX IN 2014

SUBJECT AREAS



TUS IS STRONGEST IN

CHEMISTRY

* based on WFC 2014

WFC

27.04

* based on 2014 data in the 'overall' category

TUS IS IN THE:

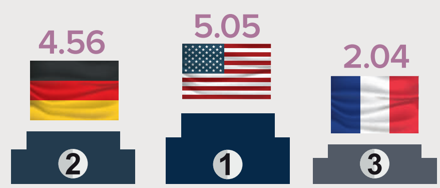
TOP 6%

of global institutions in the Nature Index*

* based on WFC in the ‘overall’ category, using 2014 data

245

Co-authored papers with 245 collaborating institutions/companies in the Nature Index



TUS collaborates most with institutions / companies in these 3 countries

* based on 2014 collaboration score

SOCIAL IMPACT

DETECTION OF URANIUM AND CHEMICAL STATE ANALYSIS OF INDIVIDUAL RADIOACTIVE MICROPARTICLES EMITTED FROM THE FUKUSHIMA NUCLEAR ACCIDENT USING MULTIPLE SYNCHROTRON RADIATION X-RAY ANALYSES

JOURNAL:
ANALYTICAL CHEMISTRY

PUBLISHED:
2014

* date obtained 9/12/2015

532

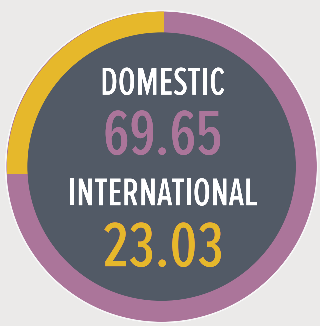
altmetric.com/details/2572954

TOP 5 INTERNATIONAL COLLABORATORS

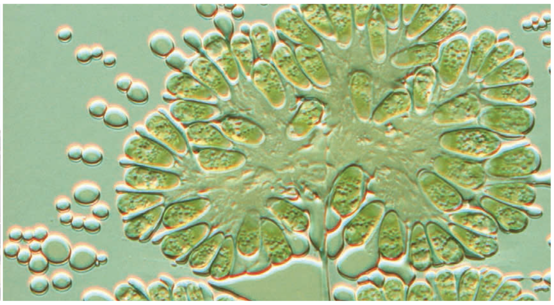
TOKYO UNIVERSITY OF SCIENCE

1. Helmholtz Association of German Research Centres, Germany [1.25]
2. Vanderbilt University (VU), United States of America (USA) [1]
3. BASF SE, Germany [1]
4. Chinese Academy of Sciences (CAS), China [0.95]
5. University of Szeged, Hungary [0.86]

WHO DOES TUS COLLABORATE WITH?



* based on 2014 collaboration score



UNIVERSITY OF TSUKUBA:

Transborder collaboration for a brighter future

As a leading national research university at the heart of Tsukuba Science City, the University of Tsukuba is spearheading transborder collaboration across all organizational, institutional and national barriers in order to promote transdisciplinary research and education to realize breakthrough knowledge and solutions.

The University of Tsukuba was founded in 1872 as modern Japan's first state higher education institution. It was reborn 42 years ago as the anchor institution of the country's premier Science City with the goal of promoting transdisciplinary education and research and being open to society. Responding to the transformation of today's society, the university is embarking on the next stage of its evolution by adopting all-encompassing reforms to become a transborder university. These reforms are based on the conviction that the complex issues of the twenty-first century require global collaboration.

Transborder education

Two of the university's most ambitious platforms are its Campus-in-Campus Initiative and Re-inventing Japan Program. The Campus-in-Campus Initiative promotes campus sharing with partners of the university, allowing students and researchers full access to global resources while promoting collaboration across national borders. Joint education programmes will be developed, and preeminent international researchers will be co-appointed.

The Re-inventing Japan Program is a government-funded scheme for developing cooperative programmes targeting specific regions. The university has been awarded funding for programmes focused on Europe, South-East Asia, Russia and Central Asia, and South America.

World-class research and innovation

The University of Tsukuba strives to create new disciplines through promoting wide-ranging collaboration in both basic science and innovation. Its Center for Computational Science has received worldwide recognition — including two Gordon Bell prizes — for its supercomputer architecture. It is promoting multidisciplinary computational science through powering research that spans particle physics, astrophysics, nuclear physics, nanoscience, life and environmental science, and information science. The Tsukuba Advance Research Alliance is seeking to unravel the mysteries of life. For example, it is exploring the molecular mechanism of lifespan and the control of immune response by regulating cellular activation and deactivation. Through the Campus-in-Campus Initiative, the university has established research units led by international principal investigators in the fields of cancer research and Buddhist philosophy.

Transborder collaboration has led to innovations that are set to spawn new industries. Cybernic technology, which integrates neuroscience, physiology, robotics, computer science, medical science and engineering in cyborg-type medical robots (HAL), is pioneering nerve regeneration using interactive biofeedback. Algal biomass development research has resulted in the discovery of two algal species with hydrocarbon-producing capabilities 10–100 times greater than higher plants, making them a highly attractive potential alternative to petroleum. The products obtained from these algae are also very promising for use in pharmaceuticals and cosmetics. In addition, the Next-Generation Social System Project seeks to promote collaboration across the academia–industry divide by realizing a sustainable mobile society in conjunction with Toyota Motor Corporation.

Legacy of global collaboration

As Japan's top sports science university, Tsukuba is supporting the 2020 Tokyo Olympic and Paralympic Games in all areas. This involvement is a legacy of Jigoro Kano, one of its founding principals and the father of modern judo. His motto "Mutual prosperity for self and others" still resonates as the university endeavours to secure a brighter future for all.

University of Tsukuba
筑波大学

Contact

1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
Phone: **+81-29-853-2111**
Web: **www.tsukuba.ac.jp/en**
Facebook: **www.facebook.com/univ.tsukuba.en**

WASEDA UNIVERSITY

A progressive spirit brings academic freedom

Located in central Tokyo, Waseda University is one of Japan's leading teaching and research universities. Since its foundation in 1882, Waseda has been preparing students for leadership in an ever-changing world. The university's innovative spirit guides the work of approximately 600,000 alumni in more than 100 countries.

Waseda University is known for its entrepreneurial character, passed through the legacy of its founder, Shigenobu Ohkuma, who emphasized the importance of maintaining an enterprising spirit and a critical mind. The university's open and diverse spirit has produced countless leaders in industry, academia and government, including seven prime ministers, and the founders and CEOs of many multinational companies, such as Sony, Samsung and Uniqlo-brand operator Fast Retailing. The university ranked 33rd in the world and the top in Japan in the graduate employability rankings released in December 2015 by Quacquarelli Symonds.

"Our aim is to create unique people, who may not fit in at conventional organizations, but who challenge orthodox thinking, do something interesting and stand out in their fields," says Waseda University president, Kaoru Kamata. "Their spirit is born out of our multicultural diversity, which Waseda is proud of," he adds.

Waseda has more than 53,000 students, over 5,000 of whom are from overseas, hailing from more than 100 countries, making Waseda the most international campus in Japan according to Japan Student Services Organization (JASSO) data. JASSO also ranks the university top in Japan for sending its students abroad for study.

A truly global university

Unlike many Japanese universities, which have only recently started accepting students from overseas, Waseda has a long tradition and established systems to help international students enjoy learning on campus. Just two years after its foundation in 1882, Waseda received its first overseas student from Korea, who was followed by the first Chinese student in 1899. The university's ties with China have been particularly strong since the late nineteenth century, as many who had studied at Waseda helped shape the Chinese Revolution of 1911, including a famous Japanese revolutionary and philosopher, Toten Miyazaki, who supported Sun Yat-sen, China's first president.

Other figures with strong ties to Waseda have also been active in creating cultural bridges between Japan and other nations. For example, Ryusaku Tsunoda, an alumnus of the inaugural class, established the Japanese Culture Center at Columbia University in 1929, globalizing Japanese studies by melding Japanese and Western perspectives and approaches.

"Since our foundation, we have looked to the world stage," says Kamata. "We want students to harness what they have learned at Waseda for the benefit of the world."

Currently, six undergraduate schools and ten graduate schools, or 2,400 courses, are taught in English. The number of programmes that allow students to obtain degrees entirely in English now stands at 50.

Waseda has also formed some 700 partnerships with higher-education institutions in 79 countries, allowing students to attend classes or even obtain degrees from other universities, including Columbia University and Peking University. The university has

also upgraded its curriculum, inviting professors from institutions across the world to offer small and interactive classes, as well as introducing a four-term calendar, rather than the calendar year starting in April used by most Japanese universities. Waseda has also launched summer sessions to give students more flexibility.

Strong research base

The university's established overseas network and systems to accept international students into its world-class courses are highly valued by the Japanese government. Waseda has been awarded state-sponsored subsidies to help accelerate its plan to become an internationally recognized research organization that can promote studies linked to real social issues such as energy shortages and ageing demographics. The university set up five research organizations to work closely with private companies, speeding up research for tackling these issues.

"Government funding is a driver helping to push through our initiatives to generate high-level research that will contribute to the world at large," Kamata says.

Moreover, Waseda is increasing investment in six research units that already have a proven global track record. Those research fields are Japanese literature and cultural studies; political science and economics; health and sports science; information and communications technology (ICT) and robotics; energy and nanomaterials; and mathematical and physical sciences.

"Our researchers used to work with their overseas counterparts on an individual basis, but the university now takes the initiative in organizing academic partnerships systematically, helping to accelerate the mobility of research personnel," says Shuji Hashimoto, professor of science and engineering, and vice president of the university. Hashimoto was instrumental in mapping out the strategy to boost Waseda's global ranking. "By promoting interdisciplinary research in these six fields, we want to be a hub institution for researchers," he adds.

Waseda's partners include Columbia University for Japanese studies, the London School of Economics for political science and economics, Loughborough University for sports science, Sant'Anna School of Advanced Studies in Pisa, Italy, for ICT and robotics, Darmstadt University of Technology for mathematics, and the University of Pisa for quantum mechanics.

Frontline research

One of Waseda's greatest strengths is its Faculty of Science and Engineering, which was created as Japan's first science and engineering department, based on Shigenobu Ohkubo's vision of fusing science with engineering.

"Students who like to experiment come to Waseda, as we offer practical education along with scientific theories," says Hashimoto, whose students enjoy creating humanoid robots under his supervision. "Students are treated as real research members, discussing issues on an equal footing."

Hashimoto also points out that many of Waseda's research projects are considered promising and novel, so that they have secured external funding easily and brought results quickly. The team of scientists at the university, for example, created the original concept of CALorimetric Electron Telescope (CALET), and was soon chosen by Japan Aerospace Exploration Agency (JAXA) as its co-developer of the telescope to study cosmic high-energy phenomena and to detect dark matter. In 2015, the CALET docked at the International Space Station and started transmitting signals back to the university's data receiver. Meanwhile, researchers at Waseda who specialize in mathematical fluid dynamics won a government-related grant, which helped the university become a front-runner in nonlinear partial differential equations and their engineering applications.

Other projects that have garnered various sources of funding include studies of internal combustion systems for next-generation vehicles, innovative electric-power management systems, technologies for an eco-friendly society, and cutting-edge nanotechnology and life science. In fact, Waseda boasted the most research grant proposals accepted by the Japanese government in 13 research fields, the highest among private universities in Japan. The university also operates a bioscience research institute in Singapore, making it the first Japanese university to have a full-fledged overseas research centre where its researchers pursue collaborative research in conjunction with world-leading scientists.

"In keeping with the founder's progressive spirit, Waseda welcomes challengers," says Kamata. "Innovation doesn't happen in isolation."

Contact

www.waseda.jp/top/en
1-104 Totsukamachi, Shinjuku-ku, Tokyo, 169-8050, Japan
+81-3-3203-7747
research-info@list.waseda.jp

早稲田大学
WASEDA University